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?

A)

The most optimal value of alpha for:

i. *Ridge*: 0.05

ii. *Lasso*: 0.001

Observations:

- Upon doubling the alpha values **Lasso** seems to drop in performance and **Ridge** seems to have the upper hand in terms of R^2 score.
- With double alpha value in **Lasso**, `KitchenQual` seems to be more important.
  Also, `BsmtFinSF1` seems to in the top 10, rather than `SaleCondition_Partial`
- In **Ridge** with double the alpha value, `OverallQual` seems to be the most important followed by `1stFlrSF` in the second place. This seems to be the opposite of what could have been seen in the case of **ridge with original alpha value**.
  Also `HouseAge` replaced `GarageCars` at the 10th position.

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?

A)

`Hyper parameters Tuned Lasso Regression` performs better and has been selected as the final model. The reasons for this decision are:

1. Even though the *train* $R2$ score is lower than that of **ridge**, the difference between *train and test* $R2$ score is less. Thus, concluding that the model is able to generalize well.
2. **Lasso regression** has the potential to reduce coefficients of insignificant variables to `0`, effectively applying **Feature Selection**. Since the data is huge, feature selection will help us to filter the most important features to report to business.

Consider all the above, **`Hyper parameters Tuned Lasso Regression`** bodes better than the other.

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?

```
A) Top 5 important features are:
1. 1stFlrSF
2. 2ndFlrSF
3. GarageCars
4. ExterQual
5. KitchenQual
```

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?

A) To make sure a model is robust and generalizable; we need to make sure of the blow conditions:

1. The metric scores should be high for better prediction, thus avoiding underfitting.

2. The metrics scores between training and testing sets must be as close to each other as possible, thus avoiding overfitting.

3. By training the model on different sample from the data we have and evaluating the scores, we can get a better idea if the model is able to estimate for all data consistently or not. In other words, applying cross-validation.

When we integrate the above principles in model building, we might have to sometimes sacrifice on an extremely high training score to get it closer to the test score. When this is done, there might sometimes be a drop in the accuracy.

By good feature engineering, hyperparameter tuning and appropriate data cleaning techniques, we can make sure that the model is generalizable and robust. Even with a sacrifice of some variance/accuracy, we might be able to fulfil the true purpose of a model, i.e. to provide reasonably well future predictions.