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?

Answer: The Optimal value of Lambda determined by **GridSearchCV** method for Ridge and Lasso models are: (Scoring parameter = R2 Score)

- **Ridge: 300**
- **Lasso: 500**

After doubling the values of lambda then making predictions, the Metrics (R2 and RMSE) have **gone down**. Please see below table for comparison:

Model	Lambda	R2 Score (test)	MSE (test)	RMSE (test)
Ridge	300 (optimal)	86.68	857069328.9321	29275.75
	600 (double)	86.04	898372035.6175	29972.85
Lasso	500 (optimal)	86.98	837431865.7070	28938.42
	1000 (double)	86.649	859129644.9202	29310.913

Top Predictor values change:

Model	Lambda	Top Predictor variables
Ridge	300 (optimal)	The top 10 features produced after doubling the value of lambda remains same only the order of few variables have changed like – *GrLivArea* is now topmost feature instead *Quality.* Also, Neighbourhood *NridgHt* is now more important that *2nd Floor* area.
	600 (double)	
Lasso	500 (optimal)	Among Top 10 features, after doubling the value of Lambda, only 1 feature is changed - HouseStyle_1Story is removed and Condition1_Norm is added.
	1000 (double)	

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?

Answer: Top features selected by both Ridge and Lasso Regression model are mostly similar. Also, the TEST scores produced by both models are also close to each other:

R2 Test Score - Ridge: 0.866
R2 Test Score - **Lasso: 0.869**

RMSE Test – Ridge: 29275.75
RMSE Test – **Lasso: 28938.42**

However, we will go with **Lasso Model** as:

- it has **High R2 score** and **Low RMSE** for Test data and
- also does Feature elimination based on coefficient values (=0)

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?

Answer:
Five most important variables initially identified by Lasso Regression were –

'OverallQual','GrLivArea','Neighborhood','BsmtExposure','GarageCars'.

After these features from source data file + performing necessary EDA, then building Lasso model, below are new top 5 new predictors:

- **2ndFlrSF**: Second floor square feet
- **1stFlrSF**: First Floor square feet
- **GarageArea**: Size of garage in square feet
- **MasVnrArea**: Masonry veneer area in square feet
- **TotalBsmtSF**: Total square feet of basement area

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?

Answer:
A model is considered to be **Robust** if the mode is stable, i.e., doesn't change drastically upon changing the data.
A model is considered **Generalizable** is the model doesn't overfit the training data and also works well with unseen or test data.
Regularization involves adding a regularization term to the cost that adds up the absolute value or the squares of the parameters of the model Therefore, when applying Regularization one must be aware of its implication on the model's Bias-Variance change..

Bias quantifies how accurate the model likely to be on test data i.e. less the Bias, more accurate the model will be.
Variance refers to the degree of changes in the model itself w.r.t. changes in the data.
One must strike balance between Bias & Variance to get a stable model. Below graph represent the relation between these –