- 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?

Ans:

The lamda value of Ridge and Lasso are 0.6 and 0.0001.

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.907257	0.906848
1	R2 Score Test	0.870878	0.870961
2	RSS Train	12.890163	12.947083
3	RSS Test	8.535675	8.530184
4	MSE Train	0.013289	0.013348
5	MSE Test	0.020518	0.020505

After doubling the value of lambda ie Ridge(1.2) and Lasso (0.0002)

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.906203	0.906848
1	R2 Score Test	0.870416	0.870961
2	RSS Train	13.036670	12.947083
3	RSS Test	8.566227	8.530184
4	MSE Train	0.013440	0.013348
5	MSE Test	0.020592	0.020505

Almost most of the values are same after we double the lambda value, however slight difference has been seen in RSS train metric where Ridge is having an edge over lasso.

Important predictor variables after the change is implemented:

- ✓ OverallQual
- ✓ MSZoning_RH
- ✓ MSZoning_RM
- ✓ GarageCars
- ✓ MSZoning_RL
- ✓ OverallCond
- ✓ SaleType_Con
- ✓ SaleCondition_Normal
- ✓ GarageType_CarPort
- ✓ Condition1_PosN

- 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?

Ans:

The lamda value of Ridge and Lasso are 0.6 and 0.0001.

Based on lambda values ridge regression does not zero any of the co efficient.

However, lasso regression zeroed 5 coefficients of features(not significant) in the selected features, thus helps in feature elimination.
Hence, I will choose Lasso regression model

- 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?

Ans:

After removing the 5 most important feature ie "MSZoning_FV", "MSZoning_RL", "MSZoning_RH", "MSZoning_RM", "GrLivArea".
We have created a new lasso model and have found the another set of 5 important feature to predict SalePrice and these are
"OverallCond",
"OverallQual",
"Building type",
"area of 2nd floor",
"House Stype",
"Building Age" and
"Shape of Property"

- 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?

Ans.

1. Difference between Test accuracy and training accuracy shall not be very wide.
2. The model shall not impacted by outliers: Outlier treatment using box plot etc is most important to get a robust model. This would help standardize the predictions made by the model.
3. Cross validation : Model shall be validated across multiple test sets and accuracy shall persist across different test sets.
4. The predicted variables shall be significant which can be determined by P-values, R2 and adjusted R2. Model shall be simple but robust too.

Implications of Accuracy of a model:

1. Get more data as possible : Having more data allows the data to train itself, instead of depending on the weak correlations and assumption.
2. Fix missing values and outliers:
If the data has missing values and outliers can lead to inaccurate model.
Outliers can affect the mean, median that we are imputing to continuous variables
3. Feature Engineering or newly derived columns :
Extract the new data from the existing data
Ex1: from “House built date” and House RemodelledDate” , we can derive a new column like “HouseAge” and “isHouseRemodelled” etc.
Ex2: From Age of the person, we can extract year, month and day and after extracting the new columns , we can drop the existing features.
4. Standardization :

- *Scaling the values*: ex: one value is in meters, the other is Kilo meters, it is important to scale these feature into one standardized unit. If we did this we can get accurate model.
- *Scaling the range* : ex: one value is in thousands, other is in between 0 to 1, then we can scale all the columns in the range of 0 to 1.

5. Feature Selection:
Based on domain knowledge, we can select important features that have significant impact on the target variable.

6. Data visualization also helps the selecting the features. Statistical parameters like p-Values, VIF can give us significant variables.

7. Right algorithm

Selecting the right machine learning algorithm is very important to get accurate model.

8. Cross validation:

Some times more accuracy will cause overfitting, we can use cross validation Technique to gain generic model accuracy.