Question 1 :- What is the optimal value of alpha for ridge & Lasso regression? What will be the change in the model if you double the value for both ridge & lasso? What will be the most important predictor variable after changing the implementation?

Answer :- 1) During the Course of Implementing the ridge & Lasso regression for Sprint Having Assignment dataset

Ridge → Alpha = 1

Lasso → Alpha = 10

when we Calculate the metrics like R2 Score, RSS & RMSE for Ridge regression

the R2 Score on training data has decreased but it has increased on testing data

when we Calculate the metrics like R2 Score, RSS, RMSE for Lasso Regression

the R2 Score of training data has decreased & it has increased in testing data

the below variables are the most important Predictor variables

- LotArea
- OverallQual
- OverallCond
- YearBuilt
- BsmtFinSF1
- TotalBsmtSF
- GrLivArea
- TotRmsAbvGrd
- Street_Pave
- RoofMatl_Metal

} Predictors are Same But the Coefficients of them Predictors has changed.

Question 2:- You have determined the optimal value of λ for ridge & lasso regression during the Assignment. Now which one will you choose to apply & why?

Answer:- R2 Score of Lasso is slightly higher than for the ~~Lasso~~ test data. So we will choose Lasso regression over the ridge regression. In order to solve the problem statement.

Question 3:- After building the model, You realised that the 5 most important predictor variable in the Lasso model are not available in the incoming data. You will now have to create another model excluding the 5 most important predictor variables. which are the 5 most important predictor variables now!

Answer:- LotArea, OverallQual, YearBuilt, BsmtFinSF1, TotalBsmtSF post dropping the variables, the R2 score built by ~~Ridge~~ Lasso decreased on the training & test data.

Now, the 5 most important variables used for prediction are

1stFlrSF
GrLivArea
Street_Pave
RoofMatl_Metal
RoofStyle_Shed.

Question 4:- How can you make sure that a model is robust & generalizable? What are the implications of the same for the accuracy of the model & why?

Answer:- The model should be generalized so that the test accuracy is not lower than the training score. The model should be accurate for the datasets other than the one which where used during training. Too much importance can't be given to outlier's in the dataset. So inturn the accuracy predicted by the model is high.

To Ensure this is not the case, the outlier's analysis needs to be done & only those which are relevant to the dataset are retained.

The outlier's which doesn't make any sense to be removed from the dataset/model.

If the Built model is not robust then. it cannot be trusted for Predictive Analysis.