Answers for Problem Statement II

1. The optimal value for ridge regression is 0.9 and lasso regression is 0.001. If we double the values it will be 1.8 and 0.002. Increasing the value means increasing the penalty making the model more generalized. The more value we increase the less the value of r2 square because more variables effect will be reduced.
   The most important variable after the changes has been implemented for ridge regression are as follows:-
   1. MSZoning_FV 2. MSZoning_RL 3. Neighborhood_Crawfor 4. MSZoning_RH 5. MSZoning_RM 6. SaleCondition_Partial 7. Neighborhood_StoneBr 8. GrLivArea 9. SaleCondition_Normal 10. Exterior1st_BrkFace
   The most important variable after the changes has been implemented for lasso regression are as follows:-
   1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. BsmtFinSF1 6. GarageArea 7. Fireplaces 8. LotArea 9. LotArea 10. LotFrontage

2. We will be using the Lasso Regression here with alpha value as 0.001. The reason to use that model is as follows
   Since the penalty is an absolute value of size, LASSO regression uses a tuning parameter called lambda of coefficients identified by cross-validation. As the lambda value increases, the lasso shrinks the coefficient tends to zero, making the variable exactly equal to 0. Lasso also does variable selection. Do a simple linear regression when the lambda value is small, and when the lambda value increases, shrinkage occurs and variables with a value of 0 are ignored by the model.

3. The five most important variables after excluding the ones not available according to this problem statement are:
   1. BsmtUnfSF
   2. HalfBath
   3. LotConfig_FR2
   4. LowQualFinSF
   5. BsmtFinType2

4. The model should be as simple as possible, at the cost of less accuracy, but more
   Robust and generalizable. It can also be understood using the bias-variance tradeoff. as simple as
   Model, more biased, but less variance and more generalizable. The implications for precision are:
   A robust and generalizable model performs equally well on both training and test data. Accuracy between training and test data does not differ much.
   Bias: Bias is an error in a model when the model is weak at learning from data. A high bias means that the model unable to retrieve data details. This model does not perform well on training and test data.

Variance: Variance is model failure when the model tries to learn a lot from the data. high variance average the model is very well trained on the training data, so it works very well on the training data, but the data was invisible to the model, so the performance is very bad when testing the data. It is important to balance bias and variance to avoid overfitting and underfitting of the data.