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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:
Optimal value of alpha for lasso = 0.01
Reason: I started with the smallest alpha increased the alpha in each iteration and found alpha = 0.01 suitable in terms of r2 and RMSE and complexity of the model
Optimal value of alpha for ridge = 2.0
Reason: Since error stabilized around 2 and to get best tradeoff between bias and variance (complexity), 2 was the best alpha. To make sure that alpha is correct, I also tried to model using alpha = 1 and alpha = 3.
Since, higher alpha means more penalization, doubling the alpha for lasso will reduce the number of predictor variable in the model while also reducing the r2. In our case, doubling alpha = 0.02 reduced the r2 to ~87 and predictor variable in the model to 14
In ridge, on doubling the alpha, there was very less change in r2 and the number of predictor variables also didn't reduce by much. From 204 -> 203. Hence, there is a wide range of selection of optimal alpha in ridge
Lasso's predictors after change (10 predictors):

1. OverallQual
2. GrLivArea
3. TotalBsmtSF
4. OverallCond
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotFrontage
9. LotArea
10. WoodDeckSF

Ridge predictors after change (10 predictors):

1. Neighborhood_Crawfor
2. MSZoning_FV
3. MSZoning_RL
4. Neighborhood_StoneBr
5. SaleCondition_Partial
6. GrLivArea
7. SaleCondition_Normal
8. OverallQual
9. MSZoning_RH
10. Condition1_Norm

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:
To choose a model, we will look at the following:

1. Bias-Variance trade off
2. Complexity
3. Robustness
4. Interpretability
5. R2 and RMSE

Looking at the above factors, lasso will be the better model, primary reason is reduction in number of pr edictor variable to ~17 predictors compared to ridge where we have ~200 predictors in the 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?

Answer:
MasVnrArea
FullBath
BsmtHalfBath
Neighborhood_Blueste
Electrical_SBrkr

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 robust when both test and train accuracy are high and close to each other. This will make the model applicable well on unseen future data. At the same time a model should be least complex (occam razor) so that it become generalizable and also easy to interpret.
Also, EDA should be properly done and following things should be kept particularly in mind: 1. Outlier tr eatment 2. Missing value imputation/treatment. 3. One should use derived metrics as well to make mor e sense of data. 4. Data cleaning should be properly done.