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?

If you double the alpha value of ridge regression and set the alpha value to 10**,** the model will try to generalize the model by applying more penalties to the curve. This simplifies the model and does not fit all the data in the record. .. From the graph, you can see that Alpha 10 causes more errors in testing and training**.**

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

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?

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. RSS should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model. This also performs variable selection and that's why i choose this to be optimal for my 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?

Those 5 most important excluded predictor variables are :-

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- GarageArea

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?

Even though a model is less accurate, it should be as simple as possible, but more robust and generalized. You can also use the Bias Variance trade-offs to understand. The simpler the model, the greater the bias, but the less the variance and the more generalized it is. In terms of accuracy, its implication is that a robust, generalizable model works equally well on both training and test data. H. Accuracy does not change significantly between training and test data.
Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.
Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.