Subjective Questions and Answers

**Q1. 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**: Lasso Optimal Value: 0.001, Ridge Optimal Value: 20
- Ridge Regression Results:
  **Optimum alpha for ridge is 20.000000**
  - R2 score (train) : 0.9168422632686922
  - R2 score (test) : 0.91946576098174
  - RMSE (train) : 0.1130842067614676
  - RMSE (test) : 0.11751139371231017

**Ridge Regression with alpha 40**
R2 score (train) : 0.9116913772068697
R2 score (test) : 0.9198880748146326
RMSE (train) : 0.11653387327738382
==========
- Lasso Regression Results:
  **Optimum lasso Regression with alpha 0.001**
  R2 score (train) : 0.9104729888281275
  R2 score (test) : 0.919305564253434
  RMSE (train) : 0.11733502455604511
  RMSE (test) : 0.11762821103629995

**lasso Regression with alpha 0.002**
R2 score (train) : 0.9017129593475985
R2 score (test) : 0.9148188329245375
RMSE (train) : 0.12294157005218179
RMSE (test) : 0.12085412838296877
As can be seen from above statistics the test r2 score is higher for optimum values of alpha

- Lasso Regression Coefficients with Optimum Alpha
  - Top Positive Coefficients
    1. GrLivArea
    2. OverallQual
    3. SaleCondition_Partial
    4. Neighborhood_Crawfor
    5. TotalBsmtSF
  - Top Negative Coefficient
    1. YearBuilt
    2. Neighborhood_IDOTRR
    3. Neighborhood_Edwards
    4. BsmtUnfSF
    5. Functional
- Ridge Regression Coefficients with Optimum Alpha

- Top Positive Coefficients
  1. GrLivArea
  2. OverallQual
  3. TotalBsmtSF
  4. Neighborhood_Crawfor
  5. SaleCondition_Normal
- Top Negative Coefficients
  1. YearBuilt
  2. Neighborhood_IDOTRR
  3. Neighborhood_Edwards
  4. Condition2_PosN
  5. Neighborhood_MeadowV

Ridge with Alpha 40

| | |
|---|---|
| GrLivArea | 0.075669 |
| OverallQual | 0.069109 |
| TotalBsmtSF | 0.054370 |
| OverallCond | 0.050924 |
| Condition1_Norm | 0.047609 |

Lasso with Alpha 0.002

| | |
|---|---|
| GrLivArea | 0.126169 |
| OverallQual | 0.081421 |
| TotalBsmtSF | 0.061143 |
| OverallCond | 0.053663 |
| Condition1_Norm | 0.049586 |

**Q2. 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**: Ridge Regression because the R2score is slightly higher for unseen data in case Ridge Regression

**Q3. 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**: The top five Features after dropping existing top significant features are:

| | |
|---|---|
| 1stFlrSF | 0.132456 |
| 2ndFlrSF | 0.110043 |

| | |
|---|---|
| Neighborhood_StoneBr | 0.095354 |
| Neighborhood_NridgHt | 0.089702 |
| Exterior1st_BrkFace | 0.077633 |

**Q4. 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**: To make a model more robust and generalizable we have to ensure that the test accuracy is no lesser than train accuracy, if incase it is very less than train accuracy which means, this is a problem of overfitting. The model should be accurate for datasets other than the ones which were used during training.

The outliers analysis should be done and only those which are relevant to the dataset needs to be retained. The features which do not make any sense shouldn't be kept. The model should not be complex meaning no unnecessary features should be present, if the features which are of no use are present then they might posses multicollinearity, Lasso regression takes care of Multi Collinearity however Ridge Doesnt