**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?

- The optimal values for Lasso and Ridge regression are 0.001 and 5 respectively. The RMSE values for the same were 0.28 and 0.27 respectively for the above alpha. Also, the Squared Value for the says it was able to explain 87.8% of the Variance in the Test Set for Lasso Regression and 88.8% Variance in the Test Set for Ridge Regression.
- On doubling the value of alpha we could notice the following : The value of the coefficients started to decrease and we had almost 136 features of a coefficient of zero for Lasso Regression. As we further increase it the coefficients will further reduce and the model will start getting biased to a few features. Similar is the case for Ridge Regression as well but the Coefficients will not completely drop down to zero.

After the Change is Implemented:

Lasso Regression:

Number of Features with Coefficient Zero

```
lassodf['Ranking'].value_counts().sort_values(ascending = False)
```

```
-0.000    136
 0.007      3
-0.007      3
 0.001      3
-0.014      2
-0.444      2
-0.005      2
 0.005      2
 0.002      2
-0.002      2
```

Value of Coefficients:

| | Feature | Ranking |
|---|---|---|
| 25 | GrLivArea | 0.429 |
| 9 | OverallQual | 0.200 |
| 75 | Neighborhood_CollgCr | 0.164 |
| 85 | Neighborhood_NoRidge | 0.139 |
| 90 | Neighborhood_SawyerW | 0.136 |
| 18 | BsmtFinSF1 | 0.133 |
| 84 | Neighborhood_NWAmes | 0.121 |
| 96 | Condition1_Feedr | 0.120 |
| 0 | constant | 0.114 |
| 38 | GarageCars | 0.098 |
| 141 | Exterior1st_BrkComm | 0.092 |
| 10 | OverallCond | 0.083 |
| 26 | BsmtFullBath | 0.079 |
| 23 | CentralAir | 0.072 |
| 111 | Condition2_RRNn | 0.068 |
| 3 | LotArea | 0.065 |
| 91 | Neighborhood_Somerst | 0.058 |
| 35 | Fireplaces | 0.055 |
| 118 | HouseStyle_1.5Unf | 0.045 |
| 43 | WoodDeckSF | 0.041 |

Ridge Regression:

Value of Coefficients:

| | Feature | Ranking |
|---|---|---|
| 25 | GrLivArea | 0.373 |
| 9 | OverallQual | 0.183 |
| 75 | Neighborhood_CollgCr | 0.154 |
| 85 | Neighborhood_NoRidge | 0.138 |
| 0 | constant | 0.123 |
| 84 | Neighborhood_NWAmes | 0.123 |
| 91 | Neighborhood_Somerst | 0.119 |
| 90 | Neighborhood_SawyerW | 0.115 |
| 141 | Exterior1st_BrkComm | 0.110 |
| 38 | GarageCars | 0.109 |
| 137 | RoofMatl_WdShake | 0.107 |
| 96 | Condition1_Feedr | 0.105 |
| 205 | SaleType_Con | 0.097 |
| 23 | CentralAir | 0.097 |
| 18 | BsmtFinSF1 | 0.096 |
| 26 | BsmtFullBath | 0.085 |
| 58 | MSZoning_RH | 0.082 |
| 10 | OverallCond | 0.080 |
| 117 | HouseStyle_1.5Fin | 0.070 |
| 118 | HouseStyle_1.5Unf | 0.069 |

Before the Change is Implemented:

Lasso Regression:

```
lassodf['Ranking'].value_counts().sort_values(ascending = False)
```

```
-0.000    122
-0.012      3
 0.003      3
 0.006      3
-0.013      2
-0.006      2
 0.080      2
 0.008      2
 0.005      2
-0.021      2
```

| | Feature | Ranking |
|---|---|---|
| **25** | GrLivArea | 0.445 |
| **75** | Neighborhood_CollgCr | 0.203 |
| **9** | OverallQual | 0.177 |
| **85** | Neighborhood_NoRidge | 0.160 |
| **91** | Neighborhood_Somerst | 0.153 |
| **90** | Neighborhood_SawyerW | 0.144 |
| **141** | Exterior1st_BrkComm | 0.139 |
| **84** | Neighborhood_NWAmes | 0.138 |
| **18** | BsmtFinSF1 | 0.138 |
| **38** | GarageCars | 0.105 |
| **96** | Condition1_Feedr | 0.101 |
| **0** | constant | 0.097 |
| **10** | OverallCond | 0.085 |
| **73** | Neighborhood_BrkSide | 0.080 |
| **26** | BsmtFullBath | 0.080 |
| **23** | CentralAir | 0.079 |
| **205** | SaleType_Con | 0.078 |
| **111** | Condition2_RRNn | 0.070 |
| **3** | LotArea | 0.069 |
| **72** | Neighborhood_BrDale | 0.067 |

Ridge Regression:

| | Feature | Ranking |
|---|---|---|
| **25** | GrLivArea | 0.394 |
| **75** | Neighborhood_CollgCr | 0.177 |
| **137** | RoofMatl_WdShake | 0.172 |
| **9** | OverallQual | 0.169 |
| **85** | Neighborhood_NoRidge | 0.162 |
| **91** | Neighborhood_Somerst | 0.160 |
| **205** | SaleType_Con | 0.153 |
| **141** | Exterior1st_BrkComm | 0.135 |
| **84** | Neighborhood_NWAmes | 0.133 |
| **90** | Neighborhood_SawyerW | 0.131 |
| **0** | constant | 0.116 |
| **106** | Condition2_Norm | 0.115 |
| **96** | Condition1_Feedr | 0.108 |
| **38** | GarageCars | 0.106 |
| **58** | MSZoning_RH | 0.104 |
| **23** | CentralAir | 0.103 |
| **57** | MSZoning_FV | 0.101 |
| **196** | GarageType_Detchd | 0.100 |
| **18** | BsmtFinSF1 | 0.098 |

**Important Features after the change is implemented is:**

**Lasso:**

1. GrLivArea
2. OverallQual
3. Neighborhood_Crawfor
4. Neighborhood_NridgHt
5. Neighborhood_Somerst

**Ridge:**

1. GrLivArea
2. OverallQual
3. Neighborhood_Crawfor
4. Neighborhood_NridgHt
5. Neighborhood_NoRid

**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?

- We will go with the Lasso Regression as both the regression techniques have provided us with almost similar performance but Lasso has helped us in choosing the features as well hence making the model somewhat simpler as compared to ridge regression.

**Question 3**

After building the model, you realized 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?

- The five most important Predictor Variables Now are:

    - RoofMatl_Membran
    - age_house
    - Neighborhood_OldTown
    - MiscFeature_Shed
    - GarageType_Attch

**Question 4**

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

- To make a robust and generalizable model, we need to keep in mind to build a model that will not have huge variation on the Model's Predictive Power for any small changes in data. This is means the model should perform well on the unseen data as well. Also try to keep the model simple so that it can generalized to varying conditions.
- Accuracy of the model depends on the above factors since a data with outliers and missing values can have a huge impact on the accuracy of the model. Hence, we need to create a model that is robust to the outliers.