Q1-Which variables are significant in predicting the price of a house?

- GrLivArea -- Above grade (ground) living area square feet, Higher the area the better is the significance
- OverallQual_VERY_GOOD --Rates the overall material and finish of the house should be 9 or 10 for better significance
- 1stFlrSF -- First Floor square feet should be higher for better significance
- Foundation_Wood - Type of foundation is wood is having negative corelation
- RoofMatl_Tar -- Roof material of type Gravel & Tar is having negative correlation

```
pd.Series(ridge_model.coef_, index = col).sort_values(ascending=False)
```

```
GrLivArea                 0.289849
OverallQual_VERY_GOOD     0.132392
1stFlrSF                  0.104851
MasVnrArea                0.075453
SaleCondition_AdjLand     0.073824
                            ...
OverallQual_VERY_POOR    -0.071594
YearBuilt_(1880, 1900]   -0.072835
LandSlope_Sev            -0.075631
RoofMatl_Tar&Grv         -0.094205
Foundation_Wood          -0.105843
Length: 125, dtype: float64
```

Q2 -- How well those variables describe the price of a house.

R2 score for ridge for train is around 91% and test is around 89 %. So we should be able to explain these price of the home to approximately 90%

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.913358	0.911640	0.731482
1	R2 Score (Test)	0.880330	0.894562	0.779642
2	MSE (Train)	0.002226	0.044395	0.077391
3	MSE (Test)	0.002858	0.050179	0.072542

SUBJECTIVE QUESTIONS

Q. 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?

Optimal Value of alpha for Ridge Regression: 0.5
Ridge Regression important predictor Variables:

```
pd.Series(ridge_model.coef_, index = col).abs().sort_values(ascending=False)
```

```
GrLivArea                  0.289849
OverallQual_VERY_GOOD      0.132392
Foundation_Wood            0.105843
1stFlrSF                   0.104851
RoofMatl_Tar&Grv           0.094205
```

Changing Optimal Value of alpha for Ridge Regression: 1
Ridge Regression important predictor Variables:

```
pd.Series(ridge_model.coef_, index = col).abs().sort_values(ascending=False)
```

```
GrLivArea                  0.269022
OverallQual_VERY_GOOD      0.130090
1stFlrSF                   0.105472
RoofMatl_Tar&Grv           0.080233
Foundation_Wood            0.077813
                             ...
```

Optimal Value of alpha for Lasso Regression: 0.005

```
pd.Series(lasso_model.coef_, index = col).abs().sort_values(ascending=False)
```

```
GrLivArea           0.169634
1stFlrSF            0.053344
ExterQual_TA        0.051320
GarageCars_3        0.049846
BsmtFinType1_GLQ    0.032867
```

Changing Optimal Value of alpha for Lasso Regression: 0.01
Lasso Regression important predictor Variables:

```
pd.Series(lasso_model.coef_, index = col).sort_values(ascending=False)
```

```
GrLivArea                  0.071130
Fireplaces_1               0.032607
BsmtFinType1_GLQ           0.029415
GarageCars_3               0.023184
Foundation_PConc           0.021996
                             ...
```

The increase in the alpha value are changing few of the predictors and co-efficient of the predictor variables

Q. 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?

I would prefer the Ridge model as the R2 Scores of train and test data are almost the same and because the values of the R2 scores are high

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.913358	0.911640	0.731482
1	R2 Score (Test)	0.880330	0.894562	0.779642
2	MSE (Train)	0.002226	0.044395	0.077391
3	MSE (Test)	0.002858	0.050179	0.072542

Q. 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?

```
pd.Series(lasso_model.coef_, index = X_lasso_ridge_train.columns).sort_values(ascending=False).head(5)
```

```
1stFlrSF            0.112649
BsmtFullBath_1      0.015411
MasVnrType_Stone    0.011772
TotRmsAbvGrd_7      0.008284
TotRmsAbvGrd_8      0.007613
dtype: float64
```

1stFlrSF - First Floor square feet
BsmtFullBath_1 - Basement full bathrooms equal 1
MasVnrType_Stone - Masonry veneer type is stone
TotRmsAbvGrd_7 - Total rooms above grade (does not include bathrooms) = 7
TotRmsAbvGrd_8 - Total rooms above grade (does not include bathrooms) = 8

	Metric	Lasso Regression
0	R2 Score (Train)	0.625679
1	R2 Score (Test)	0.662340
2	MSE (Train)	0.091375
3	MSE (Test)	0.089798

The R2 score has fallen down considerably

Q. 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?

A model to be robust , generalizable and being accurate for a linear regression model ar e as follows

- The accuracy of test and training data should be in the same norms
- The error terms should be normally distributed
- Homoscedasticity of errors, equal variance around the line
- Error terms should have constant value
- There should be no little multi-collinearity
- No Correlation between predictor variables
- Perform cross-validation sets, average them, and compute the standard deviation. This will give you the accuracy and +/-.
- Can validate using AIC (*Akaike's Information Criteria)*, BIC (*Bayesian information criteria)* for model evaluation and selection

References:

- http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/
- https://www.statsmodels.org/
- https://waterprogramming.wordpress.com/2017/02/22/dealing-with-multicollinearity-a-brief-overview-and-introduction-to-tolerant-methods/#:~:text=Lasso%20Regression,as%20a%20measure%20of%20complexity.

- https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf
- https://stats.stackexchange.com/questions/104779/why-does-ridge-regression-work-well-in-the-presence-of-multicollinearity
-