Problem Statement -1

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

Ans -

The below variables are significant in predicting the price

- LotArea---------- Lot size in square feet
- OverallQual--------Rates the overall material and finish of the house
- OverallCond-------Rates the overall condition of the house
- YearBuilt-------- ---Original construction date
- BsmtFinSF1-------Type 1 finished square feet
- TotalBsmtSF------Total square feet of basement area
- GrLivArea----------Above grade (ground) living area square feet
- TotRmsAbvGrd---Total rooms above grade (does not include bathrooms)
- Street_Pave-------Pave road access to property
- RoofMatl_Metal--Roof material_Metal

	Ridge	Lasso
LotArea	59778.431939	63955.064210
OverallQual	115599.252408	119957.483345
OverallCond	35638.745398	37354.981812
YearBuilt	54545.692314	53864.332906
BsmtFinSF1	51586.657410	50216.539701
TotalBsmtSF	76674.754264	78348.099735
1stFlrSF	73061.086063	8832.898863
2ndFlrSF	37149.879346	0.000000
GrLivArea	87839.676484	163982.920640
BedroomAbvGr	-52962.603870	-62831.358381
TotRmsAbvGrd	52937.952456	51280.023696
Street_Pave	49959.412426	63045.460825
LandSlope_Sev	-27846.862924	-37188.510825
Condition2_PosN	-11908.785655	-21920.323877
RoofStyle_Shed	11641.731102	17801.452620
RoofMatl_Metal	18201.049929	32845.684073
Exterior1st_Stone	-37132.047065	-69633.615929
Exterior2nd_CBlock	-32941.699298	-60463.906721
ExterQual_Gd	-54900.543840	-58459.152105
ExterQual_TA	-62317.508218	-64902.622534
BsmtCond_Po	-2488.039788	0.000000
KitchenQual_TA	-5437.664855	-4495.491440
Functional_Maj2	-23574.925049	-40743.007254
SaleType_CWD	-27224.575631	-35460.118834

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

Ans-

	Ridge Regression	Lasso Regression
R2 score(Train)	0.88	0.88
R2 score(Test)-	0.87	0.86

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.861162e-01	8.843400e-01	8.859222e-01
1	R2 Score (Test)	8.621985e-01	8.696133e-01	8.646666e-01
2	RSS (Train)	5.757188e+11	5.846979e+11	5.766994e+11
3	RSS (Test)	3.429000e+11	3.244493e+11	3.367584e+11
4	MSE (Train)	2.539098e+04	2.558822e+04	2.541260e+04
5	MSE (Test)	2.791627e+04	2.715483e+04	2.766514e+04

	Ridge	Lasso
LotArea	59778.4319	63955.0642
OverallQual	115599.252	119957.483
OverallCond	35638.7454	37354.9818
YearBuilt	54545.6923	53864.3329
BsmtFinSF1	51586.6574	50216.5397
TotalBsmtSF	76674.7543	78348.0997
1stFlrSF	73061.0861	8832.89886
2ndFlrSF	37149.8793	0
GrLivArea	87839.6765	163982.921
BedroomAbvGr	-52962.6039	-62831.3584
TotRmsAbvGrd	52937.9525	51280.0237
Street_Pave	49959.4124	63045.4608
LandSlope_Sev	-27846.8629	-37188.5108
Condition2_PosN	-11908.7857	-21920.3239
RoofStyle_Shed	11641.7311	17801.4526
RoofMatl_Metal	18201.0499	32845.6841
Exterior1st_Stone	-37132.0471	-69633.6159
Exterior2nd_CBlock	-32941.6993	-60463.9067
ExterQual_Gd	-54900.5438	-58459.1521
ExterQual_TA	-62317.5082	-64902.6225
BsmtCond_Po	-2488.03979	0
KitchenQual_TA	-5437.66486	-4495.49144
Functional_Maj2	-23574.925	-40743.0073
SaleType_CWD	-27224.5756	-35460.1188
SaleType_Con	21036.1938	25659.7557

Problem Statement -2

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-

- LotArea----------Lot size in square feet
- OverallQual--------Rates the overall material and finish of the house
- OverallCond-------Rates the overall condition of the house

- YearBuilt----------Original construction date
- BsmtFinSF1-------Type 1 finished square feet
- TotalBsmtSF------ Total square feet of basement area
- GrLivArea----------Above grade (ground) living area square feet
- TotRmsAbvGrd----Total rooms above grade (does not include bathrooms)
- Street_Pave--------Pave road access to property
- RoofMatl_Metal----Roof material_Metal

Predictors are same but the coefficent of these predictor has changed

	Ridge2	Ridge	Lasso	Lasso20
LotArea	52892.418502	59778.431939	63955.064210	63617.887669
OverallQual	106429.293471	115599.252408	119957.483345	121719.072148
OverallCond	30969.119664	35638.745398	37354.981812	36948.765235
YearBuilt	53872.884932	54545.692314	53864.332906	53764.548095
BsmtFinSF1	53388.964692	51586.657410	50216.539701	50458.153814
TotalBsmtSF	71811.348552	76674.754264	78348.099735	78209.333502
1stFlrSF	70196.443400	73061.086063	8832.898863	8244.958141
2ndFlrSF	33666.888170	37149.879346	0.000000	0.000000
GrLivArea	83295.309506	87839.676484	163982.920640	162804.680303
BedroomAbvGr	-38094.981167	-52962.603870	-62831.358381	-61134.170375
TotRmsAbvGrd	54102.652478	52937.952456	51280.023696	50757.774874
Street_Pave	34001.153057	49959.412426	63045.460825	59515.001052
LandSlope_Sev	-17857.132747	-27846.862924	-37188.510825	-29661.614776
Condition2_PosN	-3031.699352	-11908.785655	-21920.323877	-11645.855795
RoofStyle_Shed	5474.383816	11641.731102	17801.452620	1966.058339
RoofMatl_Metal	8130.068994	18201.049929	32845.684073	16580.031007
Exterior1st_Stone	-17057.383837	-37132.047065	-69633.615929	-59674.587283
Exterior2nd_CBlock	-15569.072249	-32941.699298	-60463.906721	-49678.514531
ExterQual_Gd	-49400.503457	-54900.543840	-58459.152105	-57016.336034
ExterQual_TA	-59179.903853	-62317.508218	-64902.622534	-63508.829030
BsmtCond_Po	-4343.870481	-2488.039788	0.000000	-0.000000
KitchenQual_TA	-7060.140437	-5437.664855	-4495.491440	-4450.468043
Functional_Maj2	-10968.231950	-23574.925049	-40743.007254	-31654.783158
SaleType_CWD	-16897.367011	-27224.575631	-35460.118834	-30830.830798

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- The r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem

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 five most important predictor variables

- 11stFlrSF - First Floor square feet
- GrLivArea - Above grade (ground) living area square feet

- Street_Pave - Pave road access to property
- RoofMatl_Metal - Roof material_Metal
- RoofStyle_Shed -Type of roof(Shed)

	Lasso21
OverallCond	7403.774043
1stFlrSF	163379.262938
2ndFlrSF	12227.759048
GrLivArea	186638.919740
BedroomAbvGr	-71218.036474
TotRmsAbvGrd	41610.305613
Street_Pave	101376.262107
LandSlope_Sev	-40205.679947
Condition2_PosN	0.000000
RoofStyle_Shed	53262.728685
RoofMatl_Metal	84219.173436
Exterior1st_Stone	-124162.644239
Exterior2nd_CBlock	-139534.253019
ExterQual_Gd	-77170.982079
ExterQual_TA	-108569.936019
BsmtCond_Po	-122646.594039
KitchenQual_TA	-11135.858324
Functional_Maj2	-48462.215856
SaleType_CWD	-64725.438438
SaleType_Con	52937.625483

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-

The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.