Question 1

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?

```
Ridge {'alpha': 100}
Lasso  {'alpha': 0.001}
```

Base model features of significance determined using the values of coefficients sorted in descending sequence (ranked most important & below),

Features	**Ridge**
GrLivArea	4.71E-02
1stFlrSF	3.52E-02
TotalBsmtSF	3.19E-02
LotArea	2.73E-02
BsmtFinSF1	2.56E-02
2ndFlrSF	2.36E-02
GarageArea	2.25E-02
OverallQual_8	2.22E-02
OverallQual_9	2.19E-02

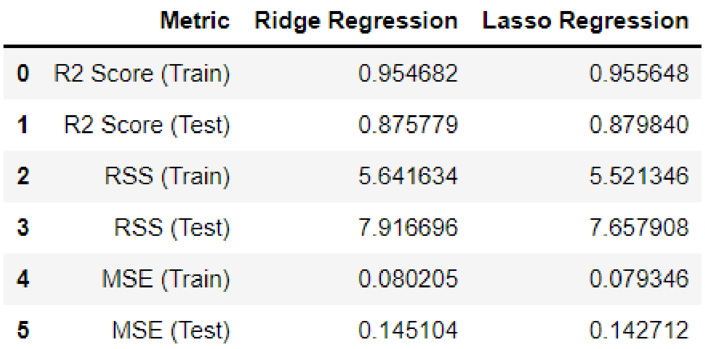
Features	**Lasso**
GrLivArea	1.31E-01
TotalBsmtSF	3.46E-02
LotArea	3.18E-02
BsmtFinSF1	2.90E-02
OverallQual_9	2.68E-02
OverallQual_8	2.63E-02
GarageArea	2.50E-02
Neighborhood_Crawfor	2.24E-02
OverallCond_8	1.94E-02
OverallCond_7	1.91E-02
Exterior1st_BrkFace	1.55E-02
BsmtExposure_Gd	1.53E-02
OverallCond_9	1.52E-02

If we double the alpha, both the models will push the coefficient values closer to zero thereby eliminating most of the features selected in the base model described above. The most important predictor will remain GrLivArea (GrLivArea: Above grade (ground) living area square feet).

Question 2

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?

Here are the R2 scores, RSS and MSE scores from both the models. As can be seen below, the scores are pretty much neck-to-neck across both train and test indicating that both the models are able to fairly generalize through a base understanding of underlying data.



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.954682	0.955648
1	R2 Score (Test)	0.875779	0.879840
2	RSS (Train)	5.641634	5.521346
3	RSS (Test)	7.916696	7.657908
4	MSE (Train)	0.080205	0.079346
5	MSE (Test)	0.145104	0.142712

However, the Lasso model has identified 107 features as insignificant as part of the feature selection process and has narrowed it down to ~190 features of importance thereby simplifying the overall model. Due to this, we will choose to implement Lasso instead of Ridge regression. Features eliminated by Lasso are below,

Features	Lasso
Utilities	0.00E+00
BsmtUnfSF	0.00E+00
1stFlrSF	0.00E+00
2ndFlrSF	0.00E+00
ScreenPorch	0.00E+00
PoolArea	0.00E+00
MiscVal	0.00E+00
YrSold	0.00E+00
MSSubClass_40	0.00E+00
MSSubClass_75	0.00E+00
MSSubClass_85	0.00E+00
MSSubClass_120	0.00E+00
MSSubClass_180	0.00E+00
MSSubClass_190	0.00E+00
MSZoning_RH	0.00E+00
MSZoning_RM	0.00E+00
LotShape_IR2	0.00E+00

LandContour_HLS	0.00E+00
LandContour_Lvl	0.00E+00
LandSlope_Sev	0.00E+00
Neighborhood_Blueste	0.00E+00
Neighborhood_BrDale	0.00E+00
Neighborhood_CollgCr	0.00E+00
Neighborhood_Gilbert	0.00E+00
Neighborhood_NAmes	0.00E+00
Neighborhood_SWISU	0.00E+00
Neighborhood_Sawyer	0.00E+00
Neighborhood_SawyerW	0.00E+00
Neighborhood_Timber	0.00E+00
Condition1_Feedr	0.00E+00
Condition1_PosA	0.00E+00
Condition2_Feedr	0.00E+00
Condition2_Norm	0.00E+00
Condition2_RRAn	0.00E+00
Condition2_RRNn	0.00E+00
BldgType_2fmCon	0.00E+00
BldgType_TwnhsE	0.00E+00
HouseStyle_1.5Unf	0.00E+00
HouseStyle_1Story	0.00E+00
HouseStyle_2Story	0.00E+00
HouseStyle_SFoyer	0.00E+00
HouseStyle_SLvl	0.00E+00
OverallQual_6	0.00E+00
OverallCond_6	0.00E+00
RoofStyle_Hip	0.00E+00
RoofStyle_Shed	0.00E+00
RoofMatl_Roll	0.00E+00
Exterior1st_AsphShn	0.00E+00
Exterior1st_CBlock	0.00E+00
Exterior1st_CemntBd	0.00E+00
Exterior1st_Stone	0.00E+00
Exterior1st_VinylSd	0.00E+00
Exterior1st_Wd Sdng	0.00E+00
Exterior2nd_AsphShn	0.00E+00
Exterior2nd_Brk Cmn	0.00E+00
Exterior2nd_CBlock	0.00E+00
Exterior2nd_CmentBd	0.00E+00
Exterior2nd_MetalSd	0.00E+00
Exterior2nd_Stone	0.00E+00
Exterior2nd_VinylSd	0.00E+00
ExterQual_Gd	0.00E+00
ExterCond_Fa	0.00E+00
ExterCond_Gd	0.00E+00

Foundation_Stone	0.00E+00
BsmtQual_Fa	0.00E+00
BsmtQual_Gd	0.00E+00
BsmtCond_Po	0.00E+00
BsmtExposure_No	0.00E+00
BsmtFinType1_BLQ	0.00E+00
BsmtFinType1_Rec	0.00E+00
BsmtFinType1_Unf	0.00E+00
BsmtFinType2_GLQ	0.00E+00
BsmtFinType2_LwQ	0.00E+00
Heating_GasA	0.00E+00
Electrical_FuseP	0.00E+00
BsmtFullBath_3	0.00E+00
BsmtHalfBath_1	0.00E+00
FullBath_1	0.00E+00
BedroomAbvGr_3	0.00E+00
BedroomAbvGr_4	0.00E+00
BedroomAbvGr_6	0.00E+00
BedroomAbvGr_8	0.00E+00
KitchenAbvGr_2	0.00E+00
KitchenAbvGr_3	0.00E+00
TotRmsAbvGrd_3	0.00E+00
TotRmsAbvGrd_9	0.00E+00
TotRmsAbvGrd_10	0.00E+00
TotRmsAbvGrd_14	0.00E+00
Functional_Maj2	0.00E+00
Functional_Min1	0.00E+00
Functional_Min2	0.00E+00
Functional_Sev	0.00E+00
Fireplaces_1	0.00E+00
FireplaceQu_Gd	0.00E+00
FireplaceQu_TA	0.00E+00
GarageType_BuiltIn	0.00E+00
GarageType_Detchd	0.00E+00
GarageCars_2	0.00E+00
GarageQual_Po	0.00E+00
GarageQual_TA	0.00E+00
GarageCond_Fa	0.00E+00
GarageCond_Gd	0.00E+00
GarageCond_Po	0.00E+00
GarageCond_TA	0.00E+00
Fence_No Fence	0.00E+00
SaleType_ConLI	0.00E+00
SaleType_WD	0.00E+00

Question 3

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?

The five most important predictors after coming to the understanding that the top 5 are no longer available in the input data are,

OverallQual_8 (Overall material and finish of the house rated as “Very Good”)

GarageArea (Size of the Garage Area in Square Feet)

Neighborhood_Crawfor (“Crawford” neighborhood which is a physical location within Ames city limits)

OverallCond_8 (Overall condition of the house rated as “Very Good”)

OverallCond_7 (Overall condition of the house rated as “Good”)

Exterior1st_BrkFace (Exterior covering on the house is “Brick Face”)

Question 4

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?

Typically underfitted models suffer from “high bias” and “low variance” whereby the model generalizes well but the accuracy on training and testing sets will not be acceptable. However, an overfitted model will have “low bias” and “high variance” thereby scoring really well on training dataset and suffers from being able to generalize. With that said, a model can be made more robust and generalisable by selecting an appropriate value of lambda (regularization) parameter. By doing so, the bias-variance trade-offs will be met whereby the model will move much closer to “low bias” and “low variance”. Overall,

- Model should not be impacted by the presence of outliers in training
- Generalizes well so that test scores remain closer to train scores with a +/- 5% margin of error
- Simpler in-terms of model complexity whereby the number of features aren’t extreme and easier to interpret the results

The accuracy of robust and generalizable models will be better than simpler/naïve models that underfit the data. However, their accuracy may not be as high as very complex models that overfit the training data. With that said, accuracy along with interpretability should be taken into consideration while presenting to prospective clients and based on the client feedback/approval the appropriate model should be deployed in production.