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 alpha value for Ridge regression is 0.3 and that for Lasso Regression is 0.0001. The R2 values for these two are given as follows:

Metric	Ridge Regression (alpha = 0.3)	Lasso Regression (alpha=0.0001)
R2 Score (Train)	0.840442	0.837936
R2 Score (Test)	0.827088	0.8351

Top 5 Predictive variables for Ridge and Lasso and their co-efficient are given below:

Feature	Ridge co-efficient
GrLivArea	0.146991
1stFlrSF	0.140458
2ndFlrSF	0.078537
LotArea	0.070077
MasVnrArea	0.061554

Feature	Lasso co-efficient
GrLivArea	0.305678
MasVnrArea	0.057972
OverallQual	0.019064
LandSlope_encoded	0.018292
BsmtFullBath	0.016104

From the above table we observe that feature GrLivArea (Living Area in sq feet) is the most preferred parameter for purchasing a house, in both methods.

If we double the alpha values, the R2 score would be as follows:

Metric	Ridge Regression (alpha = 0.6)	Lasso Regression (alpha = 0.0002)
R2 Score (Train)	0.840078	0.835703
R2 Score (Test)	0.832886	0.82927

The top 5 predictive variables with this alpha value for these two methods are:

Feature	Ridge co-efficient
GrLivArea	0.133323
1stFlrSF	0.12534
2ndFlrSF	0.075694
MasVnrArea	0.0613
LotArea	0.059772

Feature	Lasso co-efficient
GrLivArea	0.259055
MasVnrArea	0.052878
OverallQual	0.020144
LandSlope_encoded	0.018193
BsmtFullBath	0.015881

Observation: In both tests alpha values, the Ridge and Lasso predicted that the GrLivArea to be most prominent feature with highest co-efficient value.

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?

The optimal value for Ridge and Lasso are 0.3 and 0.0001 respectively that yielded the best result during our model fitting exercise. Their R2 scores are given as follows:

Metric	Ridge Regression (alpha = 0.3)	Lasso Regression (alpha=0.0001)
R2 Score (Train)	0.840442	0.837936
R2 Score (Test)	0.827088	0.8351

Both models have a good R2 score on both train and test data and even the difference is minimal. I would choose Lasso Regression method over Ridge as the difference between the train and test scores are comparatively lesser. However, it yielded about 36 features to have a 0 or < 0 (negative) co-efficient value. In case of Ridge, number of negative co-efficient values are 20 features. I would go with Lasso as the fit seems to be optimal and most important predictive variables are used (complete list given below):

Feature	Lasso co-efficient
GrLivArea	0.305678
MasVnrArea	0.057972
OverallQual	0.019064
LandSlope_encoded	0.018292
BsmtFullBath	0.016104
GarageCars	0.013619
WoodDeckSF	0.012849
LotArea	0.012541
RoofMatl_encoded	0.009516
Fireplaces	0.008958
LandContour_encoded	0.007751
FullBath	0.006182
ScreenPorch	0.005931
Street_encoded	0.005851
OverallCond	0.005435
Functional_encoded	0.005058
BsmtHalfBath	0.003532
TotRmsAbvGrd	0.003471
2ndFlrSF	0.002918
Foundation_encoded	0.002685
SaleCondition_encoded	0.001962
RoofStyle_encoded	0.001921
BsmtCond_encoded	0.001561
ExterCond_encoded	0.000929
Neighborhood_encoded	0.000749
PavedDrive_encoded	0.000577
LotConfig_encoded	0.000463
SaleType_encoded	0.000347
BsmtFinType2_encoded	0.000249
YearBuilt	0.000177
Heating_encoded	0.000129
Electrical_encoded	0.000108
GarageType_encoded	0.000108

YearRemodAdd	0.000096
Exterior1st_encoded	0.000013
LotFrontage	0
BsmtFinSF1	0
BsmtFinSF2	0
BsmtUnfSF	0
TotalBsmtSF	0
1stFlrSF	0
HalfBath	0
GarageArea	0
EnclosedPorch	0
3SsnPorch	0
PoolArea	0
MiscVal	0
Condition1_encoded	0
CentralAir_encoded	0
GarageQual_encoded	0
GarageCond_encoded	0
MoSold	-0.00033
MSZoning_encoded	-0.00037
HeatingQC_encoded	-0.00061
Exterior2nd_encoded	-0.00094
LotShape_encoded	-0.00136
BedroomAbvGr	-0.00153
BsmtFinType1_encoded	-0.00205
YrSold	-0.00208
BldgType_encoded	-0.00221
MSSubClass_encoded	-0.0023
GarageFinish_encoded	-0.00312
HouseStyle_encoded	-0.00336
OpenPorchSF	-0.00388
BsmtExposure_encoded	-0.00661
BsmtQual_encoded	-0.00751
ExterQual_encoded	-0.00923
KitchenAbvGr	-0.00949
Condition2_encoded	-0.00978
LowQualFinSF	-0.01147
KitchenQual_encoded	-0.01316

Question 3

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?

After removing the 5 most important variables in Lasso Model - GrLivArea, MasVnrArea, OverallQual, LandSlope_encoded and BsmtFullBath, we re-did the Lasso modelling. The R-square values for train and test sets are: 0.8185 and 0.7515 respectively.

Now the new five most important predictor variables are:

Feature	Lasso Co-efficient
1stFlrSF	0.28977
2ndFlrSF	0.156374
TotalBsmtSF	0.07803
BsmtFinSF1	0.043055
LotArea	0.040324

Question 4

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?

To ensure that the model is robust and generalisable the rule of thumb are applied – that is, the model should not overfit or underfit while applying it on the unseen data. In our examples, we got the train and test values as close as possible for the optimal value of alpha parameter after applying regularization techniques. Further even when the alpha was doubled, the results we got were almost similar – though the R2 values are different, the difference between train and test data sets were minimal. This definitely seem to indicate that model is accurate.