Assignment-based Subjective Questions

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 value of alpha for ridge is 1 and lasso regression is 10

   After double the value of of alpha for both models, the changes of RMSE and r-squared has changed slightly different, which are as follow:

	Before double alpha	After double alpha
Ridge	RMSE: Train set: 24640.845515845547 Test set: 27789.05534365291 R-squared: Train set: 0.8757553693463557 Test set: 0.8457228669687125	RMSE: Train set: 24855.405107769282 Test set: 27736.71541290967 R-squared: Train set: 0.8735822346088256 Test set: 0.8463034732522918
Lasso	RMSE: Train set: 24503.03524410003 Test set: 27876.98857653622 R-squared: Train set: 0.8771412232680478 Test set: 0.844744960338683	RMSE: Train set: 24597.685987879842 Test set: 27804.16497557916 R-squared: Train set: 0.8761902280788325 Test set: 0.8455550523814387

   Ridge: RMSE is lower and r-squared is higher for test set after doubling alpha
   Lasso: RMSE is lower and r-squared is higher for test set after doubling alpha
   The most important predictors after the change stay the same for ridge but slightly different for lasso:

	Before change	After change
Ridge	GrLivArea, OverallQual, GarageArea, Neighborhood_NoRidge, TotalBsmtSF	GrLivArea, OverallQual, GarageArea, Neighborhood_NoRidge, TotalBsmtSF
Lasso	GrLivArea, OverallQual, GarageArea, Neighborhood_NoRidge, TotalBsmtSF	GrLivArea, OverallQual, GarageArea, TotalBsmtSF, Neighborhood_NoRidge

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 regression model of choice would be Ridge regression model with alpha value of 1. The reason is that the RMSE value is the lowest and the r-squared is the highest for the test set. Full information is as follow:

Model	RMSE Train	RMSE Test	r2 Train	r2 Test
ridge	24640.845516	27789.055344	0.875755	0.845723
lasso	24503.036471	27876.992890	0.877141	0.844745

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 predictors variables now?
   The five most important predictors after rebuilding lasso are:

Top 5 features	**Coef values**
TotRmsAbvGrd	102057.5284
GarageCars	78459.63564
1stFlrSF	75400.01429
BsmtQual	69072.05177
Neighborhood_MeadowV	-54683.09122

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 the model is robust and generalizable, there are some approaches:

- Cross validation and hold-out strategy to evaluate model performance on different subsets.
- Apply the regularisation on the regression model
- Different values of alpha for ridge or lasso technique
- Different scoring approach such as negative mean squared errors or others

The implications for the accuracy of the model would depends on the trade-off between bias and variance. High bias and low variance would result in underfitting and vice versa. Therefore these problems can be avoided by using the methods above to achieve a model with low bias and variance.