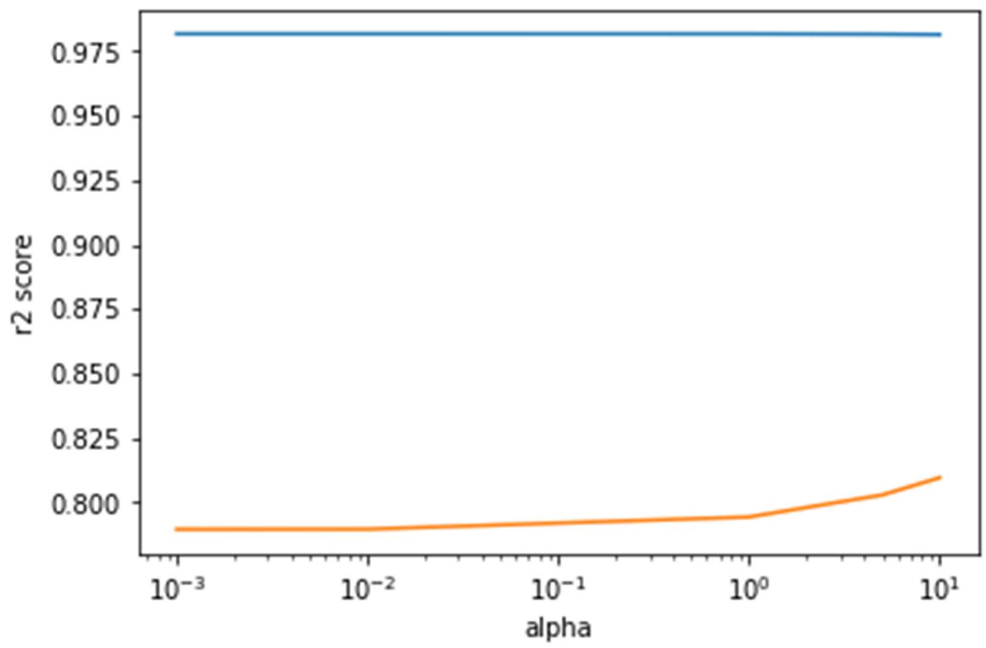
Question 1

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?

Ridge and lasso are two types of Regularization techniques. Regularization helps to solve the problem of overfitting. The optimal value of alpha for ridge regression is 15 and for lasso is 600. The alpha which is the hyperparameter of the model which manages the trade-off by adjusting the penalty. If hyperparameter is doubled model will be more complex to the cost function. Hence higher the lambda value will bias the model. The most important predictor variables are: “MSSubclass,” “LotFrontage,” “LotArea.”

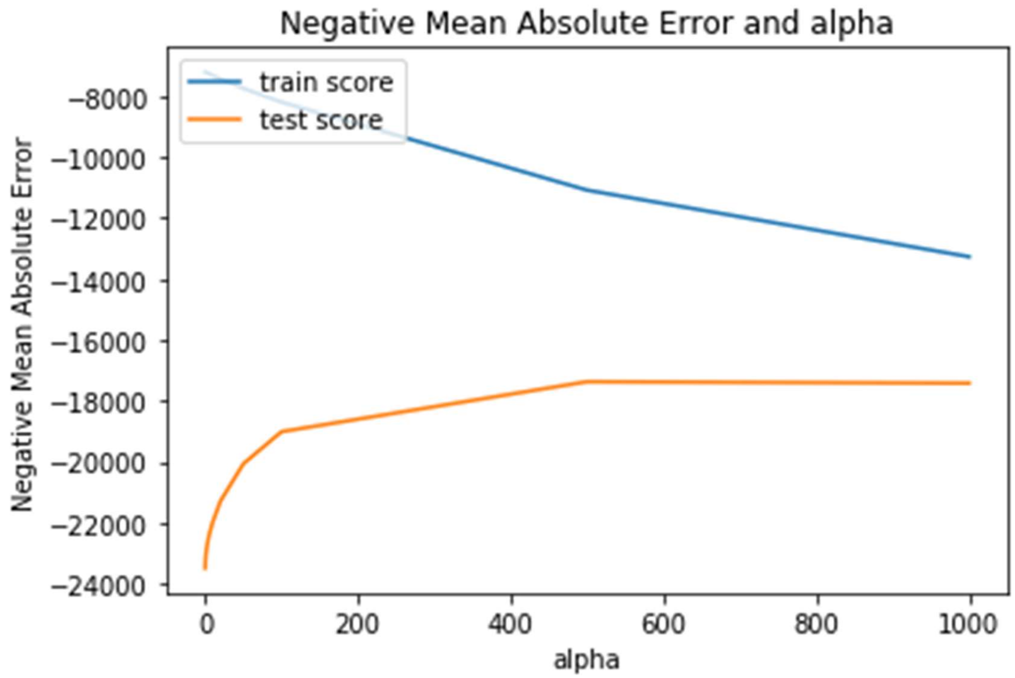
Lasso trying different values of alpha:

```
# lasso regression

lm = Lasso(alpha=300)

Score:.96

Top Features

'LotFrontage', 0.0),

 ('LotArea', 6579.670367266532),

 ('OverallQual', 9755.126237521872),

 ('YearBuilt', 9377.996335180207),

 ('YearRemodAdd', 2705.7679465251367),
```



alpha =600

Score: 96



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 two most commonly used regularised regression methods: ridge regression and lasso regression. Both these methods are used to make a regression model simpler while balancing the 'bias-variance' trade-off. In ridge regression, an additional term, 'sum of the squares of the coefficients', is added to the cost function along with the error term, whereas in case of lasso regression, a regularisation term, 'sum of the absolute value of the coefficients', is added.

The optimal value for ridge and lasso alpha is 15 and 600. one of the most important benefits of lasso regression is that it results in model parameters, such that the lesser important features' coefficients, becoming zero. In other words, lasso regression indirectly performs feature selection. Hence performing lasso regression will be chosen to build the model since Lasso trims down the coefficients of redundant variables to zero and, thus, indirectly performs variable selection as well. Ridge, on the other hand, reduces the coefficients to arbitrarily low values, though not zero.

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?

```
'LotFrontage', -100.76015170645766),
 ('LotArea', 7494.43399855977),
 ('OverallQual', 6961.407004364695),
 ('YearBuilt', 10891.275950174844),
 ('YearRemodAdd', 2210.858530398313),
```

Question 4

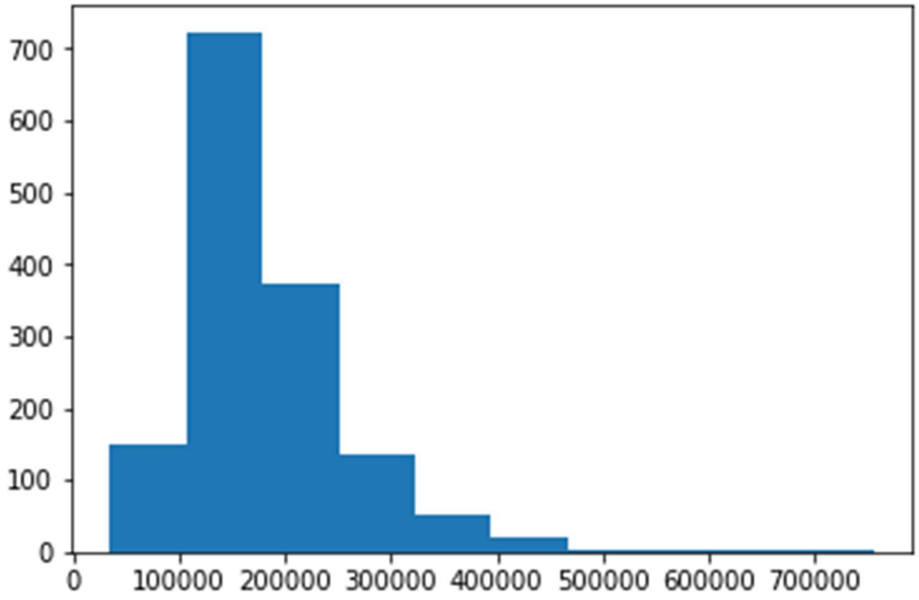
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Making the model more robust and generalisable requires it to remove overfitting and underfitting. Overfitting occurs when the model fits the training data too well and do not perform well in an unseen data. Overfitting occurs when the model shows low bias and high variance. It results in complicated model. In other case underfitting occurs when the model cannot capture the data. In other words when the model cannot fit the data well. Which results in simple model. These problems can be overcome with using validation or cross validation.

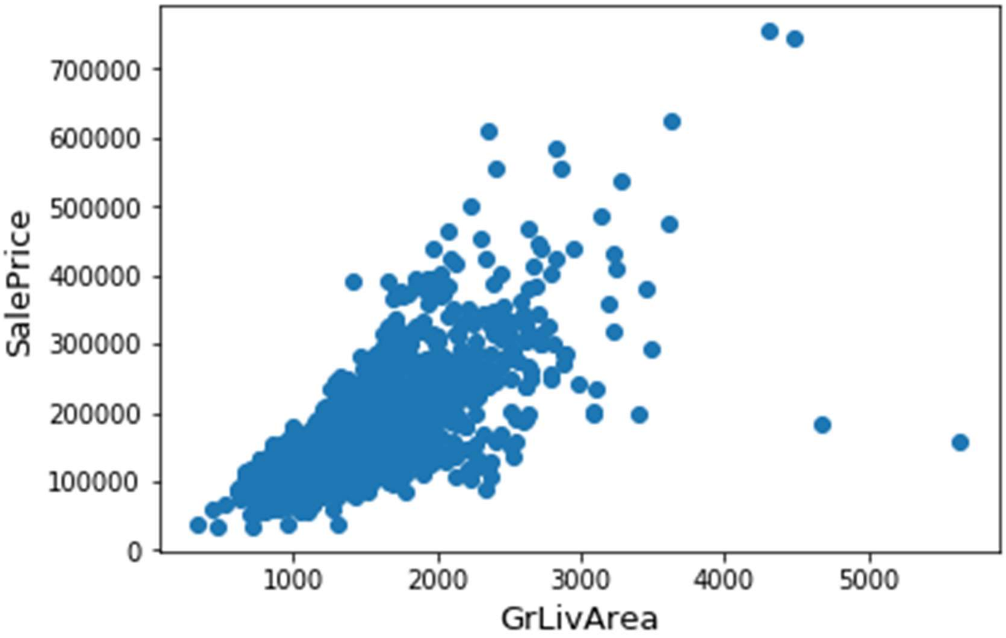
Part 1

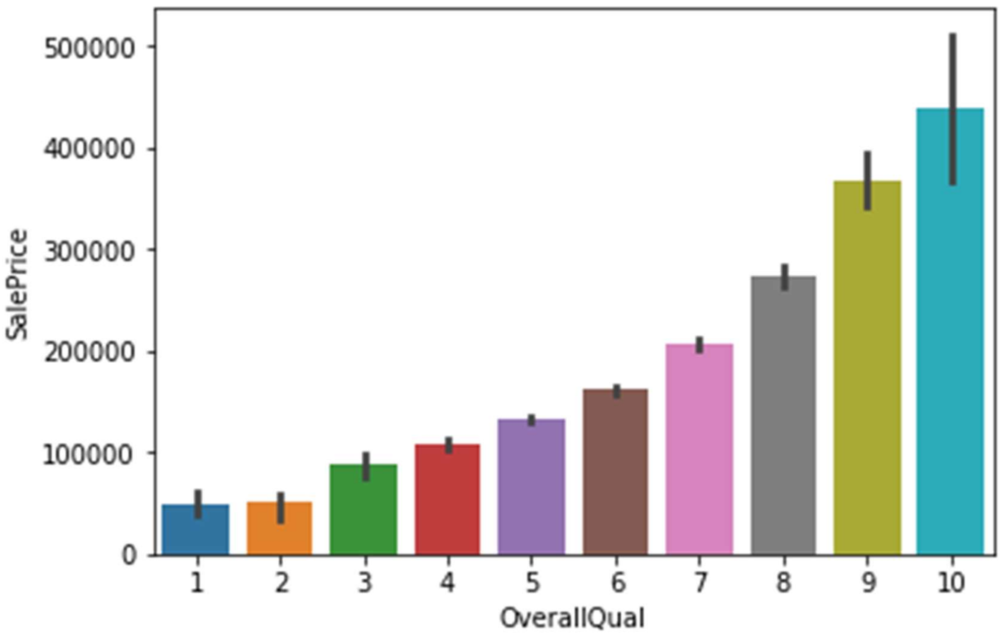
- Which variables are significant in predicting the price of a house, and
- How well those variables describe the price of a house.

There are many factors that have an impact in the target variable “Sale Price” which are quality of the house, overall size of the house, Lot front Age, Year built. The price of the house was also right skewed

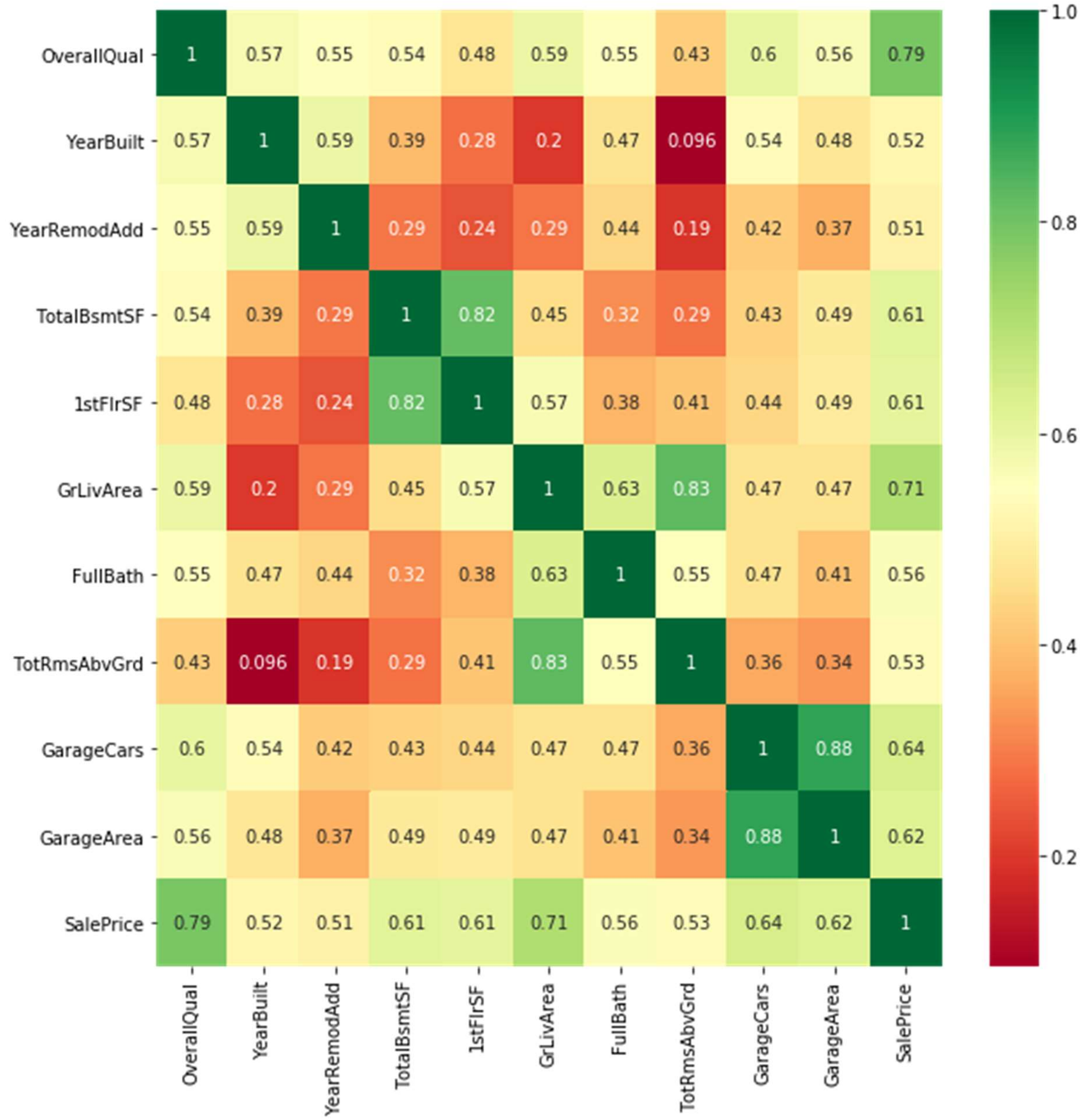


Also GrLivArea and Overallquall feature is correlated with the Sale Price.





The most correlated features are:



	OverallQual	YearBuilt	YearRemodAdd	TotalBsmtSF	1stFlrSF	GrLivArea	FullBath	TotRmsAbvGrd	GarageCars	GarageArea	SalePrice
OverallQual	1	0.57	0.55	0.54	0.48	0.59	0.55	0.43	0.6	0.56	0.79
YearBuilt	0.57	1	0.59	0.39	0.28	0.2	0.47	0.096	0.54	0.48	0.52
YearRemodAdd	0.55	0.59	1	0.29	0.24	0.29	0.44	0.19	0.42	0.37	0.51
TotalBsmtSF	0.54	0.39	0.29	1	0.82	0.45	0.32	0.29	0.43	0.49	0.61
1stFlrSF	0.48	0.28	0.24	0.82	1	0.57	0.38	0.41	0.44	0.49	0.61
GrLivArea	0.59	0.2	0.29	0.45	0.57	1	0.63	0.83	0.47	0.47	0.71
FullBath	0.55	0.47	0.44	0.32	0.38	0.63	1	0.55	0.47	0.41	0.56
TotRmsAbvGrd	0.43	0.096	0.19	0.29	0.41	0.83	0.55	1	0.36	0.34	0.53
GarageCars	0.6	0.54	0.42	0.43	0.44	0.47	0.47	0.36	1	0.88	0.64
GarageArea	0.56	0.48	0.37	0.49	0.49	0.47	0.41	0.34	0.88	1	0.62
SalePrice	0.79	0.52	0.51	0.61	0.61	0.71	0.56	0.53	0.64	0.62	1

OverallQual, GrLivArea, GarageCars.