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?

Answer: The Linear Model I created didn't face overfitting problem. The model provided high accuracy and low rmse. The cause for the model performance is that I used a business and Data Analysis process to handle categorical variables instead of the Automated Dummies creation.

Yet I tried Ridge and Lasso regression just for theoretical prove of the accuracy of the linear model. Here are the parameters for the Linear, Ridge, and Lasso models. The Optimal Values of Alpha for both Ridge and Lasso were near 0 (0.001); Refer to the Jupyter NoteBook submitted for more details and all Values of Alpha, R2, and RSME:

Linear Model:

```
              r2                  rsme
Train:   0.8558424348307      5.143167133927194
Test:    0.8726894033325      4.779025969790417
```

Ridge:

```
Alpha =  0.001
Train:   0.8739614130597315   4.838036374984554
Test:    0.8333528264020111   4.838036374984554
--------------------------------------------------
Alpha =  0.01
Train:   0.8739613129205608   4.838038296923169
Test:    0.8333711044495397   4.838038296923169
--------------------------------------------------
Alpha =  0.1
Train:   0.8739517126696268   4.838222547879538
Test:    0.8335389331900581   4.838222547879538
```

Lasso:

```
Alpha =  0.001
Train:   0.8739442256468397   4.838366236244295
Test:    0.833488411607123    4.838366236244295
--------------------------------------------------
Alpha =  0.01
Train:   0.8726757778568227   4.862648555370103
Test:    0.833333914214835    4.862648555370103
--------------------------------------------------
Alpha =  0.1
Train:   0.8477794459647319   5.31684519615497
Test:    0.8082468716955079   5.31684519615497
```

As mentioned the linear model didn't face overfitting problem. For testing purposes, the increase in Alpha for both Ridge and Lasso leads to decrease in R2 and increase in rsme.

The most important variables were identified by the model were: GrLivArea (23.439658); OverallQual (20.732023); Neighborhood has good influence on price (NridgHt,StoneBr, and NoRidge) are the most expensive; OverallCond, LotArea, Garage Areas and TotalBsmtSF, and total rooms are also significant. Refer to the notebook for full details of all variables and their coefficients.

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?

Answer: My Linear model has no overfitting problem, so there was no need for Ridge or Lasso regression. Just for theoretical discussion, The best value was near Zero for both Ridge and Lasso (0.001). For the model I will use it will be the linear model (neither Ridge, nor Lasso).

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?

Answer: Based on the calculated Coefficient in the model the variables significance are defined as follows. The next 5 most important are highlighted.

	var	coef
7	GrLivArea	24.087006
1	OverallQual	18.26868
2	OverallCond	9.63765
19	Neighborhood_NridgHt	8.81698
56	GarageScore	8.536302
30	Neighborhood_StoneBr	7.862806
4	BsmtFinSF1	7.753747
0	LotArea	6.490234
35	Neighborhood_Veenker	6.391886
6	TotalBsmtSF	5.017376
26	Neighborhood_NoRidge	4.629694
53	RoomsScore	3.280434
39	HouseStyle_Split	2.638501
17	Neighborhood_Somerst	2.582968
10	OpenPorchSF	2.279766

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?

Answer: The model is robust and generalisable as its R2 and rmse on the test data is as good as it is on the train data. To ensure this I have tested the model on many different splits of train/test and all provided almost the same excellent results with high r2 and low rmse. The residual analysis also proves the robustness and generalisability of the model.