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?

As per the assignment solved, optimal alpha = 500

R2 value on train and test data is similar

```
0.8814085789649619  → R2 train
0.894360234303091 -> R2 test
724474799664.773
267352365367.40036
724474799.664773
623198986.8703971
```

The RSS increases with increase in alpha and model complexity reduces.

If we double the value of alpha, then for Ridge regression we get the following metrics

```
0.8615627513090742 → R2 train
0.8826702291249962 → R2 test
845712928778.117
296937347072.94275
845712928.778117
692161648.1886778

Significant higher value of alpha cause underfitting.
```

Similarly for Lasso when alpha is doubled

```
0.8936017477200449
0.8907197767997861
649986751423.0073
276565609244.9049
649986751.4230074
644675079.8249532
```

Alpha=500

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.020279e-01	8.814086e-01	8.990729e-01
1	R2 Score (Test)	8.881393e-01	8.943602e-01	8.926741e-01
2	RSS (Train)	5.985113e+11	7.244748e+11	6.165636e+11
3	RSS (Test)	2.830962e+11	2.673524e+11	2.716195e+11
4	MSE (Train)	2.446449e+04	2.691607e+04	2.483070e+04
5	MSE (Test)	2.568848e+04	2.496395e+04	2.516239e+04

Alpha=1000

Metric	Linear Regression	Ridge Regression	Lasso Regression	
0	R2 Score (Train)	9.020279e-01	8.615628e-01	8.936017e-01
1	R2 Score (Test)	8.881393e-01	8.826702e-01	8.907198e-01
2	RSS (Train)	5.985113e+11	8.457129e+11	6.499868e+11
3	RSS (Test)	2.830962e+11	2.969373e+11	2.765656e+11
4	MSE (Train)	2.446449e+04	2.908114e+04	2.549484e+04
5	MSE (Test)	2.568848e+04	2.630897e+04	2.539045e+04

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?

Lasso Regression will be the right choice as it eliminates features and results in more robust model

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?

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?

Model is robust and generalized if it performs well on unseen or test data. To ensure this, simple model should be preferred over complex model.

Simple model with less number of features will not cause overfitting and have higher accuracy on test data. Remove features that have high correlation. Data cleaning and outlier pruning of data will generate model that give higher accuracy.