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 Optimum vales for alpha in Lasso and Ridge as per the model I have developed are mentioned below:

1. Ridge – 2
2. Lasso – 0.01

If we increase the values to below:

1. Ridge – 4
2. Lasso – 0.02

On increasing the alpha value for LASSO from 0.01 to 0.02, the R2 value decreases and RMSE increases drastically. Also you can notice that the variables with non-zero coefficients have changed. Earlier with 0.01 we had 20 such variables but now with the doubled alpha value we can see that this number has reduced to 14. The newly selection predictor variables are mentioned below.

	Variable	Coeff
0	constant	12.016
4	OverallQual	0.166
24	GarageCars	0.161
22	Fireplaces	0.134
33	MSZoning_RL	0.048
165	BsmtFinType1_GLQ	0.031
151	Foundation_PConc	0.017
38	LotShape_Reg	-0.002
99	RoofStyle_Gable	-0.007
163	BsmtExposure_No	-0.009
141	MasVnrType_None	-0.027
157	BsmtQual_TA	-0.078
190	KitchenQual_TA	-0.082
145	ExterQual_TA	-0.104
203	GarageFinish_Unf	-0.141

Whereas in RIDGE, despite doubling the alpha value, there is no significant difference noticed in the model. With respect to the predictor variables, the no of such variables still remain constant at 220. Most important ones being the below (in ascending order)

	Variable	Coeff
0	constant	10.763
4	OverallQual	0.407
14	GrLivArea	0.236
5	OverallCond	0.234
24	GarageCars	0.230
...	...	...
157	BsmtQual_TA	-0.109
176	Heating_Grav	-0.139
58	Neighborhood_MeadowV	-0.150
191	Functional_Maj2	-0.162
84	Condition2_PosN	-0.215

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:
Despite the R2 value and RMSE being better for RIDGE compared to LASSO, I will make a decision to go with LASSO due to the simplicity of the model. The RIDGE process resulted in us getting a count of 220 variables with co-efficient being non zero, whereas the LASSO process resulted in just 21. Hence, LASSO model would be much more simpler with an acceptable R2 value of above 75 percent.

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:
After creating a new model, the new top 5 predictor variables with LASSO process and alpha value as 0.01 are:
OverallQual - Rates the overall material and finish of the house
GarageCars - Size of garage in car capacity
GrLivArea - Above grade (ground) living area square feet
FullBath - Full bathrooms above grade
BsmtExposure_Gd - Refers to walkout or garden level walls (Good Exposure)

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:
A model is said to be robust if any variation in the data does not impact the performance much and result in bad predictions. A generalised model would typically adopt to new unseen data samples from the same source. Overfitting is something we need to make sure to avoid, to keep the robustness of the model intact. In the assignment with the Surprise Housing, we chose to go in for the LASSO model over RIDGE as it is more robust and generalised. To put it in other words, a model should not be complex in nature. The more the complexity, the less the robustness and less the generalization behaviour of the model.

Now coming from the perspective of ACCURACY, a highly complex model will definitely have high accuracy (like what we witnessed with RIDGE in our use case). In order to make our model robust and generalised, we would have to decrease the variance which will lead to some bias, which means accuracy will decrease. This is a trade-off, we need to make to have a really robust and generalised model.