Subjective Questions and Answers

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?

- The optimal value of alpha for ridge was 0.3 and for lasso was 0.0001.
- For Ridge , If we double the value from 0.3 to 0.6 , we can see that coefficients value gets reduced which indicated more regularization of the model
- For Lasso , if we double the value from 0.0001 to 0.0002, similar to ridge was the observation more the value of alphas coefficient values get reduced more the regularization

Below is the comparison screenshot:

```
#alpha = 0.6
ridge = Ridge(alpha=0.3)
ridge.fit(X_train_ridge, y_train)
coefficient_sum_3 = sum(list(ridge.coef_))
ridge = Ridge(alpha=0.6)
ridge.fit(X_train_ridge, y_train)
coefficient_sum_6 = sum(list(ridge.coef_))
```

```
print("Difference in ridge coefficients",(coefficient_sum_3- coefficient_sum_6) )
```

```
Difference 0.00034611358052538677
```

```
lasso = Lasso(alpha=0.0001)
lasso.fit(X_train_lasso, y_train)
lasso.coef_
coefficient_sum_1 = sum(list(lasso.coef_))
lasso = Lasso(alpha=0.0002)
lasso.fit(X_train_lasso, y_train)
lasso.coef_
coefficient_sum_2 = sum(list(lasso.coef_))
```

```
print("Difference in lasso coefficients",(coefficient_sum_1- coefficient_sum_2) )
```

```
Difference in lasso coefficients 0.0006387080563243774
```

The most important predictor variables after change is implemented as:

- GrLivArea - Above grade (ground) living area square feet , the higher the value more the price
- YearBuilt - the newer the house the higher the price we get for the house
- OverallQual - better the quality of the house fetches better prices
- BsmtFinSF1 - Type 1 finished square feet , positively correlated
- KitchenQual - Kitchen quality better the pricier the house

- OverallCond - Better the overall condition of the house, the higher rates of the house
- TotalBsmtSF - Total square feet of basement area, more the basement area more pricy the house becomes
- LotArea - Higher the lot area more the price value of the house

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 determined values of alpha for ridge is 0.3 and lasso is 0.0001. We will choose lasso regression as it:

- Reduces the coefficients as well as it reduces the constant to zero.
- Plus when we have large number of features lasso attempts to reduce the number of features, so it helps in reducing the number of features
- Plus regularized models have smooth curves

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?

The five most important variables in the lasso model are GrLivArea, YearBuilt, OverallQual, BsmtFinSF1, KitchenQual.

After excluding the above, most important variables are:

- OverallCond - Rates the overall condition of the house, the better the condition of the house , the better the prices of the house
- TotalBsmtSF - Total square feet of basement area, higher the square feet the more expensive the house becomes
- LotArea - Lot size in square feet, the more the area available the pricier the house
- Neighborhood_Crawfor - Physical locations within Ames city limits : Crawford seems to be an expensive neighbourhood to live in
- GarageArea - Size of garage in square feet, the more the size of the garage , the higher the asset value of the house

Question 4: How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

We can make sure that a model is robust and generalizable by ensuring the following

- By reducing the number of features

- Reducing multi-collinearity of the model by avoiding features that are highly correlated
- Checking whether the features are significant or not by analysing the p-value
- By reducing the variance
- Making the model simple and giving explain ability of features in the model significance

Accuracy of the model might reduce slightly if we make the model more generalizable in the training set but it will outperform in the test cases most of the time as it has generalized the model and given importance to features that matter.

So as a general concept when model is more generalised or made more simple , the accuracy in the training set might decrease but the accuracy in the unknown dataset or test dataset will improve in most of the cases.

The reason behind this behaviour can be attributed to the model giving significance to few features that matter the most, and the model being relatively simple can explain the reason why it is predicting such behaviour because of the driver variables