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?

In Ridge Regression, the optimal value of alpha is 200, while in Lasso Regression, it’s 500.

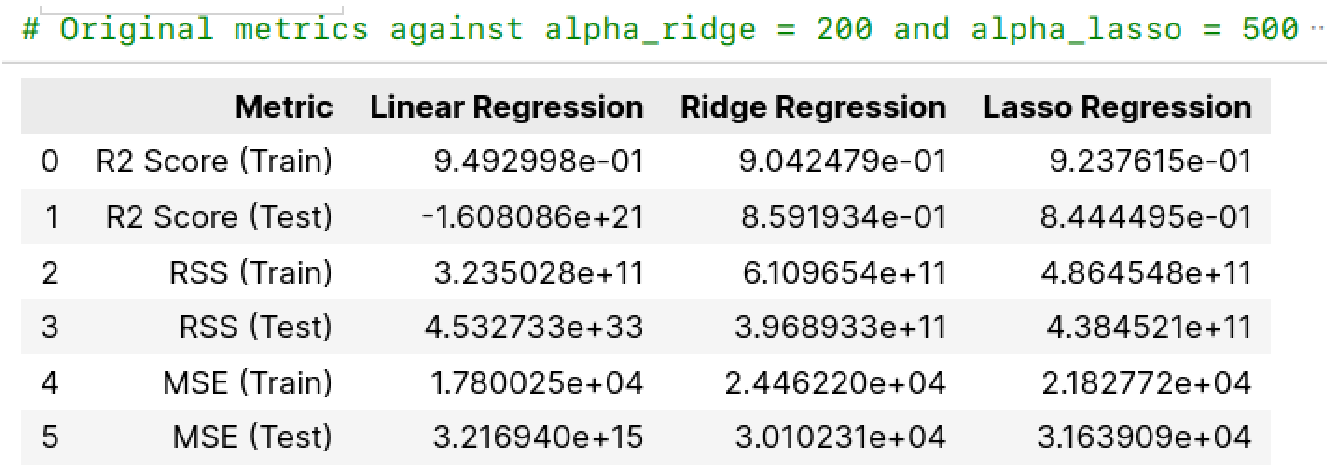
With doubling alpha:

- R2 score for train set reduced in both regression models which is not a good indicator, which brings it closer to the R2 score of test sets.
- MSE has seen slight increase in both models.
- In Ridge, `OverallQual` and `GrLivArea` switched top two places when alpha was doubled.
- `Neighborhood_NoRidge` came 3rd after being 5th before doubling alpha, followed by `Neighborhood_NridgHt`, which indicates that the location contributes a lot in the price.
- `1stFlrSF` has taken over `2ndFlrSF`, indicating that price is mostly decided by the main floor area.
- In Lasso, all roof materials types were dropped from top features.
- `GrLivArea` came 1st after being 2nd before doubling alpha, followed by `OverallQual`.
- `lasso` and `ridge` has predicted the same top 4 variables, meaning that the new (doubled) alpha values made both models the same.
- `GarageCars` came in the fifth place. It was one of the highest correlated values to `SalePrice` before running regression.

Most important variables after doubling alpha are:

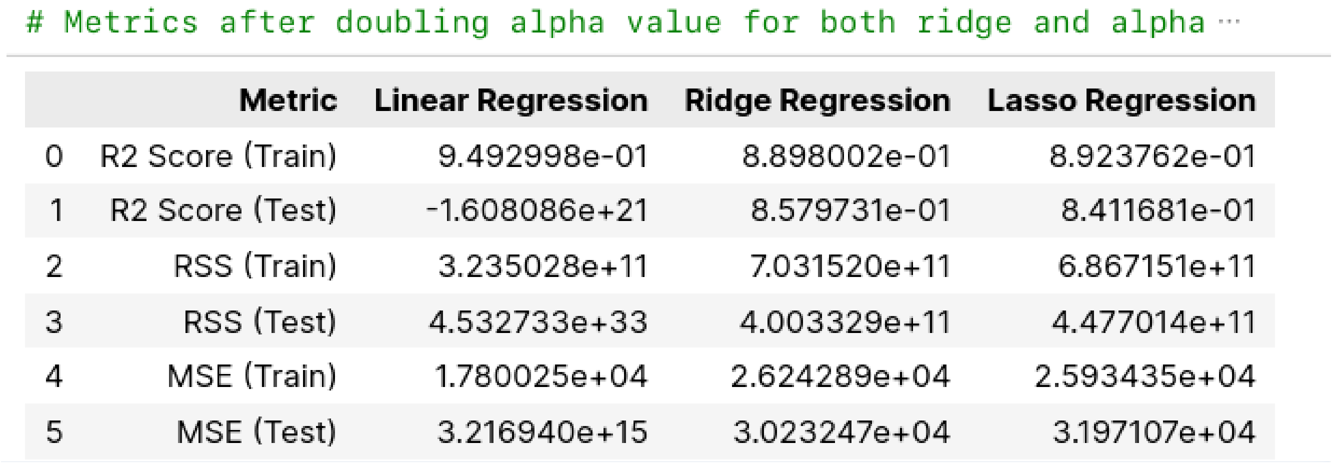
1. OverallQual 9146.38
2. GrLivArea 8691.37
3. Neighborhood_NoRidge 6361.30
4. Neighborhood_NridgHt 5674.340
5. 1stFlrSF 5618.97

Figure (1): Metrics for Linear, Ridge and Lasso regressions for Train and Test sets.



```
# Original metrics against alpha_ridge = 200 and alpha_lasso = 500 ···
```

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.492998e-01	9.042479e-01	9.237615e-01
1	R2 Score (Test)	-1.608086e+21	8.591934e-01	8.444495e-01
2	RSS (Train)	3.235028e+11	6.109654e+11	4.864548e+11
3	RSS (Test)	4.532733e+33	3.968933e+11	4.384521e+11
4	MSE (Train)	1.780025e+04	2.446220e+04	2.182772e+04
5	MSE (Test)	3.216940e+15	3.010231e+04	3.163909e+04



```
# Metrics after doubling alpha value for both ridge and alpha ···
```

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.492998e-01	8.898002e-01	8.923762e-01
1	R2 Score (Test)	-1.608086e+21	8.579731e-01	8.411681e-01
2	RSS (Train)	3.235028e+11	7.031520e+11	6.867151e+11
3	RSS (Test)	4.532733e+33	4.003329e+11	4.477014e+11
4	MSE (Train)	1.780025e+04	2.624289e+04	2.593435e+04
5	MSE (Test)	3.216940e+15	3.023247e+04	3.197107e+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?

- Ridge regression has higher R2 score, and lower gap between train and test sets.
- Doubling alpha lowered R2 score by average of 2-3% and increased MSE in both models, which is not helpful.
- Lasso's features selection is useful in case of large amount of variables. Doubling alpha also showed how Lasso has performed by dropping less features.
- On the other hand, with lower MSE on Test set in Ridge `3.01` vs `3.16` in Lasso, this implies that Ridge performed better on unseen data.

```
Lasso is preferred in this case study due to its ability to drop features
that are less significant to `SalePrice`, and this is effective since we
have a lot of variables (246) in this case.
```

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?

```
Original top 5 features:
1. RoofMatl_CompShg 32722.70
2. GrLivArea 29737.47
3. RoofMatl_WdShngl 20224.20
4. RoofMatl_Tar&Grv 20030.87
5. RoofMatl_WdShake 13766.13

New top 5 features now are:
1. 2ndFlrSF 24211.24
2. 1stFlrSF 18551.56
3. OverallQual 15063.32
4. GarageCars 7856.51
5. Neighborhood_NoRidge 7737.31
```

Question 4

How can you make sure that a model is robust and generalisable?

```
- When the model performs well on test data compared to an equivalent
training data, it means that it's generalised well. This means also the
model did not over or under fit the data.
- On the other hand, for a model to be robust, it should perform well on
test (unseen) data when training data is noisy (e.g.: missing/null
values), or has outliers or variations.
```

What are the implications of the same for the accuracy of the model and why?

```
- Robustness (as defined above) goes opposite to accuracy, as for the
model to be robust, it tends to be more complex, leading to lower bias,
higher variance, and hence too much accuracy.
- There's always a tradeoff between bias and variance. To reach an
optimal model, we'll sacrifice some robustness towards optimally
```

predicting values that were not used to train the model, which is the measurement of the accuracy.

Note:

- In this case study, we cleaned, prepared and scaled the data before splitting into training and test sets, so that both are equivalent. So our model can easily be considered to be well generalised. Below values of accuracy proves this point.
- If we wanted to make the model more robust, we could've split the data before cleaning, then train the model. In this case, the test set would include lots of noise (missing values). Then apply regularisation and testing different alpha values for both ridge and lasso.