## 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 1

In our analysis, the optimal value of alpha (lambda) was found to be 10 for Ridge regression and 0.001 for Lasso
regression.

If we double the value of alpha for Ridge regression, the penalty for larger coefficients will become more severe,
potentially causing the coefficients of less important variables to shrink towards zero, but not reaching zero. This may
slightly decrease the model's accuracy, particularly if some variables that were important at the original alpha level
become less influential at the new alpha level.

In contrast, doubling the value of alpha for Lasso regression (from 0.001 to 0.002) will still maintain the feature
selection property of the Lasso model. However, more coefficients may shrink to zero, possibly losing some important
variables and decreasing the model's accuracy.

The most important predictor variables after these changes would need to be reevaluated by running the models with the
new alpha values and examining the coefficients. However, in general, we might expect the most robust predictors from
the original model to remain important.

## 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 2

The choice between Ridge and Lasso depends on the specific situation and goals of the analysis. In your case, the Lasso regression model was suggested because it performed comparably to the Ridge model and had the added benefit of feature selection. Lasso regression could simplify the model by reducing the number of features, potentially making it easier to interpret and less prone to overfitting.

## 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 3

If the top five predictors in the original Lasso model are not available, you would need to rebuild the model excluding these variables. The new set of most important predictor variables would be identified by fitting the Lasso regression model on this updated dataset and examining the coefficients. Since this is a hypothetical scenario, I can't provide the actual variables without running the model on the updated dataset.

## 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 4

Ensuring a model is robust and generalizable involves several strategies:

- **Cross-validation**: This technique involves splitting the dataset into 'k' subsets and training the model 'k' times, each time leaving out one of the subsets from training and using it as a test set. The average error across all 'k' trials is computed. This helps to ensure the model's robustness and generalizability.

- **Train/Test Split**: By partitioning the data into a training set to build the model and a testing set to evaluate its performance, we can assess how well the model generalizes to unseen data.
- **Regularization**: Techniques like Ridge and Lasso regression can prevent overfitting by adding a penalty term to the loss function that constrains the coefficients of the model.
- **Feature Selection**: This helps to remove irrelevant or redundant predictors that could cause the model to overfit to the training data.
- **Outliers detection and treatment**: Outliers can significantly impact the performance of a model. Detecting and treating outliers can help improve model robustness.

Implications for accuracy: If a model is robust and generalizable, it should have similar accuracy when applied to both training data and new unseen data. If a model performs well on the training data but poorly on new data (high variance), it may be overfitting. Conversely, if a model performs poorly on both training data and new data (high bias), it may be underfitting. The goal is to find the right trade-off between bias and variance that achieves the most accurate predictions on new unseen data.