Question 1

Consider the linear regression model $y = X\beta + \epsilon$, where y is the dependent variable, X is the design matrix (with dimensions $n \times p$), β is the vector of coefficients, and ϵ is the error term. When applying regularization (penalization), we modify the objective function to include a penalty on the size of the coefficients. For ridge regression, this penalized form of the least-squares objective is given by:

$$\min_{\beta} \left(\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

where λ is the regularization parameter controlling the strength of the penalty.

1. Derive the closed-form solution for β (the coefficient estimates) for ridge regression.
2. Show that with regularization ($\lambda > 0$), the estimate for β is always invertible, even if $X'X$ is not full rank.

Question 2

For this exercise, you will apply ridge and lasso regression to a real dataset. We will use the popular **California Housing** dataset, which is widely available online. Write a python script to perform the following:

1. **Download the dataset:** You can access the Boston Housing dataset from the following sources:
   - `sklearn.datasets` includes the data we need under $fetch\_california\_housing$ (Note: you can find the dataset from alternative sources such as UCI or Kaggle.)
   - Alternatively, search for the dataset online and download it in CSV format.
2. **Apply Ridge and Lasso Regression:**
   - Split the data into a training set and a test set.
   - Standardize the features before fitting the model.
   - Fit ridge and lasso regression models using Python's `sklearn` library.
   - Use cross-validation to tune the hyperparameter λ (also referred to as α in `sklearn`).
3. **Compare the performance of Ridge and Lasso Regression:**
   - Report the mean squared error (MSE) on both the training and test sets.
   - Compare the coefficients obtained from both methods.