Texas Christian University
COSC 40023 Spring 2025
Assignment 3

Due: February 19th, Wednesday, at 11:30 PM (Late submission NOT accepted)
Submission (two files, no compression): `assignment3.py` and `assignment3.R` through TCU Online

Download the dataset called "Crime_Data.csv". In the dataset, each observation is corresponding to a small US city. Here are the descriptions of the columns.

`X1` = reported violent crime rate per 100,000 residents
`X2` = annual police funding in $/resident
`X3` = % of people 25 years+ with 4 years of high school
`X4` = % of 16 to 19 year-olds not in high school and not high school graduates
`X5` = % of 18 to 24 year-olds in college
`X6` = % of people 25 years+ with at least 4 years of college
`Y` = total overall reported crime rate per 1 million residents

Build multiple linear regression models in both Python and R. Two files named `assignment3.py` and `assignment3.R` should be created. Here are some additional requirements.

1. `10%` of the data should go to the test set. In addition, `random_state` must be set to `0` in Python and `seed` must be set to `123` in R.
2. Build the optimal model using backward elimination. Your program should have the intermediate code and comments that demonstrate how you found the best team of independent variables. Significance level is `5%`.
3. Test set results should be predicted through the optimal model and should be stored in `y_pred`. `y_pred` can be scaled or unscaled.
4. Print out all the coefficients and the intercept of the optimal model.
5. Print out the predicted `Y` based on the optimal model when `X1` = `500`, `X2` = `50`, `X3` = `40`, `X4` = `30`, `X5` = `20` and `X6` = `10`. There is no need to use all `X` values, and `Y` can be scaled or unscaled.
6. Have sufficient single-line and multi-line comments.