ASSIGNMENT QUESTIONS - 30 MARKS

Instructions:

- Generate the output code using R Markdown (PDF format) and save the file as (Roll. No**.rmd**)
- Write the analysis in the code itself using the # comment in R
- Read the description of the dataset required to be installed before answering the questions

Question 1

- Load the dataset **Seatbelts** from **datasets** package in R.
- Subdivide the dataset into two - Before Legislation and After Legislation. Obtain the boxplots for the **Drivers Killed Before Legislation** and **Drivers Killed After Legislation**
- Analyse and compare both the plots obtained
- Plot the boxplot of each variable in one single plot and give a short analysis of each variable
- **Aim:**
  To predict the **Drivers Killed Before Legislation** using a multi-linear regression model with 6 predictor variables (exclude Drivers and Law variable) using **Ridge and Lasso regression**. First perform the following for Ridge and then for Lasso. Use **glmnet** package
  - Use the k-fold cross-validation for **glmnet** to obtain the best lambda value (Ridge and Lasso). Plot and analyse the cross-validation curve. Set the nlambda=100 and lambda.min.ratio=0.0001
  - Once the best lambda value is obtained, obtain the coefficients of the Ridge using **predict()** and Lasso regression using **coef()**

Question 2

- Load the dataset **Swiss** from **datasets** package in R
- In one single plot, plot the boxplot of each of the given variables
- Produce a grid of scatter plots to analyse the correlation (positive or negative) between all pairs of variables in the dataset
- **Aim:**
  To build a linear regression model that predicts **Infant.Mortality** on other variables. Use **leaps** package
  - Use subset selection to select the best subset model with the most correlated variables with the response. Write the best subset of predictors (variables) chosen for one to five predictor models
  - Determine the best overall model by computing the adjusted R-squared and CP values. Mention the variables

- Finally, extract the coefficients of the best model selected using the adjusted R-squared and CP metrics

Question 3

- Load the dataset **Default** from **ISLR** package in R
  Note: Before starting the analysis, do not forget to set a random seed
- Fit a logistic regression model that uses **income** and **balance** to predict **default**
- Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
  - Split the sample set into a training set and a validation set
  - Fit a multiple logistic regression model using only the training observations.
  - Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5
  - Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
- Now consider a logistic regression model that predicts the probability of default using **income, balance**, and a dummy variable for **student** Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for **student** leads to a reduction in the test error rate.