Advanced Statistics Assignment
(40 marks)

Instructions: Follow the pattern of Practice Assignment. Include all details of R codes and outputs in the Appendix. In the main body of the assignment summary results, observations, discussions and final conclusions are to be given.

Install the R package MASS and load the package. It includes a data set named Boston containing housing values in the suburbs of Boston and a number of predictors determining the housing values. The details of all the variables are found in R.

After installing the package use the following commands in R to view the data and learn about the attributes.

```
> library(MASS)
> View(Boston)
> help(Boston)
```

Note that except for ***chas*** and ***rad***, all other variables are continuous. We will exclude ***chas*** from the analysis but include ***rad***.

Explanation of the variables is provided at the end.

Problem definition: The goal of this assignment is to predict median housing value (***medv***) based on the available attributes.

1. Do an exploratory data analysis on Boston data (use all variables except ***chas***) and report the results. What are the main observations?
2. Perform a PCA on Boston data excluding the variables ***chas*** and ***medv*** . What do you see?
3. Extract factors from Boston data (excluding the variables ***chas*** and ***medv***) with *varimax* rotation. Remember to scale the data. What do you see? Is it possible to name the factors according to what they represent?
4. Predict ***medv*** using all predictors (except ***chas)***. Develop the model on 450 observations chosen randomly. Use the rest to validate the model. Does this give good prediction?
5. Repeat the same procedure with the extracted factors. Use the same training and validation sets as above for comparison. Which prediction procedure do you recommend?

Sample code: For determination of training and test data fix a seed. It can be any integer number. Draw a random sample of 450 rows from Boston data.

```
set.seed(integer)
indexes = sample(1:nrow(Boston), size=450)
training = Boston[indexes,]
test = Boston[-indexes,]
```

Once the factors are extracted, they may be saved in a data set along with all relevant variables in Boston. If the results of factor analysis are saved in Boston.fit

```
BostonNew <- cbind(Boston, Boston.fit$scores)
```

Explanation of the variables in the data:

Crim: per capita crime rate by town.

Zn: proportion of residential land zoned for lots over 25,000 sq.ft.

Indus: proportion of non-retail business acres per town.

Chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

Nox: nitrogen oxides concentration (parts per 10 million).

Rm: average number of rooms per dwelling.

Age: proportion of owner-occupied units built prior to 1940.

Dis: weighted mean of distances to five Boston employment centres.

Rad: index of accessibility to radial highways.

Tax: full-value property-tax rate per \$10,000.

Ptratio: pupil-teacher ratio by town.

Black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

Lstat: lower status of the population (percent).

Medv: median value of owner-occupied homes in $1000s.